

Print Edition

Monday 28 September 2026

EDITION NO. 67 · 1 PIECE · 8 PAGES

IN THIS EDITION

1. The Quest for Embedded Evaluators · Don't Worry About the Vase 1
-

The Quest for Embedded Evaluators

BY ZVI MOWSHOWITZ · DON'T WORRY ABOUT THE VASE · 27 SEPTEMBER 2026

Dario Amodei's essay *We Must Pace the Frontier* committed Anthropic to embedded evaluators, who would be placed inside Anthropic and given employee-level access, so they could provide outside perspective and also reports on what was happening.

There is only one problem. Who will be the evaluators?

OpenAI followed suit on committing to the evaluators, and also issued a milquetoast but welcome call for international coordination. I will cover that here as well.

What I won't cover today, but hope to cover tomorrow, is the latest torrent of new AI hacking incidents that came to light over the weekend, which highlights that we badly need at least embedded evaluators, and plausibly far harsher measures.

For now, you need to know that there were a lot more incidents that OpenAI did not disclose, and also a new incident at OpenAI that just happened that forced them to again pause their most advanced model. I'll get right on sorting all that out.

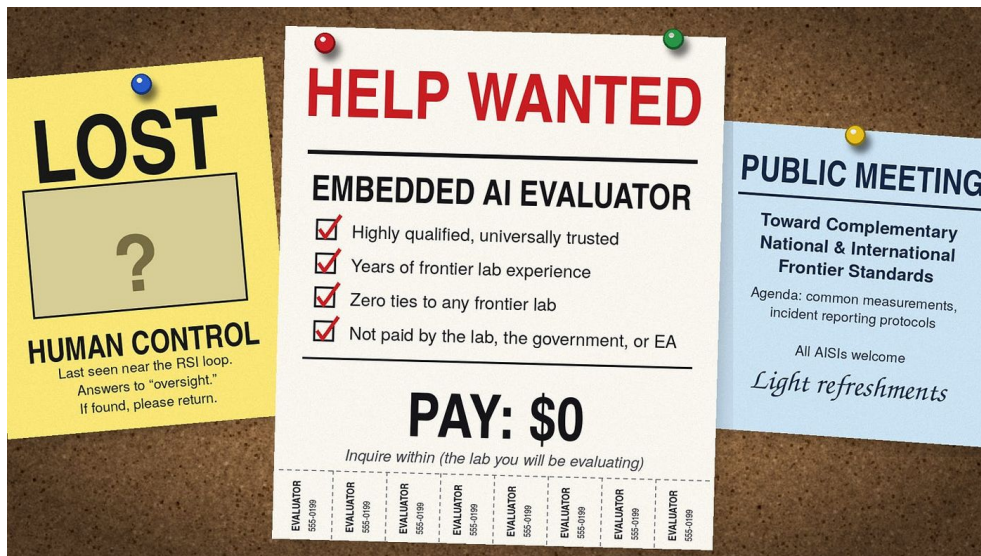


Table of Contents

1. Look, All I'm Asking For Is That You Find A Highly-Qualified, Experienced, Trustworthy, Non-Conflicted Source of Embedded Evaluators That Will Work Entirely For Free, Without Government Assistance or Money from EA Sources Not Chosen By the Lab.
2. Anthropic Partners with Accenture for Embedded Evaluation, also Plans to Include METR.
3. Reading the METR.
4. OpenAI Suggests Doing The Least We Can Do.

Look, All I'm Asking For Is That You Find A Highly-Qualified, Experienced, Trustworthy, Non-Conflicted Source of Embedded Evaluators That Will Work Entirely For Free, Without Government Assistance or Money from EA Sources Not Chosen By the Lab

Why are you telling me that this is hard?

Obviously all of the requests in the section title are individually and collectively desirable. They are all nice-to-haves. The problem is, you obviously cannot get that close to having all of them at once under present conditions.

1. Someone, somewhere, has to hire the people and pay the bill.
2. The qualified, experienced and trustworthy people need to have gained that experience and trust in some way, which is going to involve the top labs.

I am not saying the complaints are invalid, but even well-established, highly trusted auditing schemes mostly lack good solutions. The revolving door and the source of funding are hard problems. You still need to audit.

A group led by Geoffrey Hinton, Stuart Russell and Arvind Narayanan lays out their minimum standards for credible embedded evaluators in a public letter:

1. Frontier AI companies should rely on evaluators that are meaningfully independent.
 1. Payments cannot be contingent on findings, there should be no ownership or governing or other commercial relation, and no editorial control. Any conflicts of interest should be disclosed.
2. Frontier AI companies should incorporate differing viewpoints and areas of expertise.
3. Embedded evaluators should be transparent.
4. Embedded evaluators should be shielded from retaliation from the companies they embed with.
5. Frontier AI companies should grant embedded evaluators access equivalent to that of their own highly privileged employees, with exceptions for protecting data.

I agree. That seems like a good list.

Gabriel Weil pointed out in July that you do not want to let the AI developer hire their own referees. But who else is going to pick and hire them, if the government wants nothing to do with paying them and if anything is looking rather hostile to the whole thing?

We live in the stupid timeline. Weil worried that government will struggle to oversee the IVOs that would do the audits. He had not considered that the White House might actively get angry over the audits, on principle. Weil's proposal was mandatory liability insurance, which would be a good idea although the version of it that did the full job would be impossible to buy, and the version that got sold would not cover the kinds of events we most worry about.

If Anthropic is worth \$2.2 trillion and we are worried they might be 'judgment-proof' then who exactly is not judgment proof?

I would be totally up for requiring mundane (or prosaic) insurance for AI models above some capability threshold, including prior to internal deployment or weights release, not only closed external deployment. It would be marginally helpful. What it would not do is directly solve the most important problems.

One possibility is that you require buying let's say \$100 billion or \$1 trillion in insurance, not because that fixes the incentives, but because then we can require that they publish exactly what that insurance cost them, and for the insurance company to disclose the associated risk factors they used to price it.

Anthropic Partners with Accenture for Embedded Evaluation, also Plans to Include METR

They announced a partnership with Accenture.

Anthropic: The partnership will be led by Faculty, Accenture's specialist AI business, and will include evaluating and red-teaming

models, conducting alignment assessments, and testing model safeguards. Accenture helps businesses and governments deploy AI across many industries. Their understanding of how enterprises use AI in practice informs their safety approach, and they will bring that perspective to evaluating our models.

Anthropic and Accenture each expect to invest at least \$1 billion in building capacity in this area over the next five years.

... Given the importance and urgency of this work, Anthropic will fund Accenture's work directly. We are also in dialogue with METR and other nonprofit evaluators to pilot elements of embedded evaluation using their own funding. Ultimately, we believe frontier AI needs an ecosystem of evaluators operating with shared standards. We expect frontier labs to work with several organizations at once. Our partnership is non-exclusive; Anthropic will work with other evaluators to be announced in the coming weeks, and Accenture will work with other AI developers in similar capacities.

Charles: Seems totally fine to have Accenture do this to satisfy worries about METR's neutrality, as long as they also let METR do it.

From background discussions and other information, plus the plain reading here, my understanding is that Anthropic plans to use a combination of both approaches.

1. Nonprofits that come with their own funding and expertise, and can provide more of an expert inside view combined with a focus on safety concerns.
2. The big, formal, legible corporate approach that takes more of an outside view the way you would if you were a magistrate.

Drake Thomas argues for similar logic:

Drake Thomas (Anthropic): It is good for AI companies to have oversight from deeply cracked teams of experts with world-leading experience in investigating frontier alignment problems.

It is also good for AI companies to have oversight from unimpeachably boring and independent sources that even a bad-faith psyop would struggle to find complaints with.

As of September 2026, you can't max out both of these axes in one source yet, but nothing stops you from having multiple sources of independent evaluation! I think Accenture and METR are both on the pareto frontier of options. Obviously any frontier AI company which doesn't also deeply involve METR/Apollo-shaped orgs in evaluating their practices would be failing to do an adequate job of holding themselves accountable to the public (at least until such time as there exist more ordinary consultants with a deep bench of AI auditing talent, which I am hopeful will start happening soon).

And I do mean it about the pareto frontier thing, I think a lot of people are underrating Accenture? Like, go scroll through the twit-

ter account of Accenture CTO and Faculty CEO @MarcWarner10 , these are not people who haven't heard of AI safety! Upon reflection I think [the above] is a bit too one sided and want to add two caveats:

(1) Providing funding obviously introduces a COI here that there is not with eg METR.

(2) In some ways more sources of oversight are additive - additional shots on goal for noticing and being transparent about problems are great - but there's a cost where labs can emphasize the most friendly reviews or use them as a defense against more critical external review.

I have some worry that even with a cluster of very good talent doing the work, it'll be harder to say very blunt/weird critical things like "this company is taking on lots of existential risk right now and should immediately stop" from within a large "normal" company.

Scott Alexander: As a nominative determinist, I support you picking someone with the last name "Warner" for this job.

The obvious problems for using Accenture include that Accenture and Anthropic have previously partnered, and that Anthropic will be funding this operation. Neither is ideal, but again what is the alternative? This should be the job of the Faculty subdivision, which if it keeps its culture and stays independent is a reasonable option. They are in a number of ways legible and credible, and they do bring a track record.

Leo Gao (OpenAI): oh man i'm sure that accenture will have lots of expertise at, uh, *checks notes* redteaming models, assessing alignment, and testing safeguards. after all, their website talks about embracing the power of change to create 360° value and shared success for clients, people, shareh

Sy: to be fair, Faculty did have significant expertise in the evals stuff they were doing. They were working with basically all customers that matter.

Anka Reuel: It would be helpful for @Accenture to share an example of a report evaluating a frontier model. Many of my colleagues, myself included, are concerned about whether they have sufficient expertise because we haven't seen public examples demonstrating it.

@MarcWarner10

@AnthropicAI

Luke Muehlhauser is optimistic and points to some good people involved. Oliver Habryka and Dave Kasten are skeptical.

If I had to pick one evaluator to embed, I would choose METR.

If I had to pick two evaluators to embed, I would choose METR, and then I would choose someone more legible and boring. The options are not great.

In theory my first pick for the legible slot would be UK AISI, but there are obvious reasons why that presumably wouldn't play with the White

House. CAISI would be great in theory if the White House was on board and wanted to fund that. Then you look at the Big Four auditing firms, but none have the expertise and three of them are Claude shops like Accenture. Maybe MITRE or RAND, but the funding situations there have their own issues. There weren't great options.

You could perhaps help solve this problem by starting your own auditing firm today.

Reading the METR

Drake Thomas ranks the takes on embedded third party evaluators from worst to best. This taxonomy and its order seem correct to me.

The vast majority of such takes by volume are far below the zero point, and no takes make an actual case for 'no we should not have embedded evaluators.'

A ranking of takes on embedded third party evaluation this past week, from worst to best:

[contentless sneer against outgroup]

METR bad because [huge Sankey diagram]. fact checking? timelines of relative investments that make causal sense? determining which numbers are large or small fractions of other numbers? sounds like some EA bullshit to me. just look at this diagram with all those curved lines, you can SEE the nest of snakes.

METR bad because [an actual specific chain of actors who have some pairwise relationship to each other that ends with someone it would be bad for them to have strong COIs with, e.g. "METR once received funding from an org who once received funding from a person who once gave funding to a now-frontier AI company"]. It is beneath my dignity to explain how this actually affects the decisionmaking of METR; you should vaguely perform a mood affiliation and keep scrolling. Don't think too hard.

guys guys guys you HAVE to embed my company/organization into the labs. I have been an unwavering supporter of external embedded auditors since 7 minutes ago when I saw this essay taking off. Look at this thing we published once that sort of looks like AI safety if you squint! Fear not, I can assure you that I have never done anything altruistic in my life and if I had I would have been ineffective at it.

I observe that [politicized actor] has had [bad take]. Let me use this correct observation to score points for my side and politicize the situation further.

=== zero point: takes beyond this line are better than logging off ===

Dear [person with insane take], here is an earnest explanation of

why you are wrong.

METR is bad/problematic because [an actually reasonable concern, like greater cultural overlap with Anthropic than OAI or the pressure for individual employees to be on good enough terms with labs that they could later get hired], with no further suggestions for what to do about this issue.

Tweets which simultaneously acknowledge that (1) more social independence from labs would be good and (2) almost everyone competent is socially connected to labs, even if they don't propose solutions.

[literally any take that engages with object level assessment of AI companies done by a third party org like METR, SecureBio, Guidelight, Redwood, Nightingale, Apollo, etc]

=== current discourse frontier: takes beyond this line are better than anyone has yet posted ===

I have [actually reasonable concern] with METR/Redwood/etc. To address this concern, we should do [concrete proposal that makes any sense and would solve the problem].

I am founding a new third party evaluation org / pivoting my existing org to do more third party evaluation. We're doing [thing which is both (1) not identical to METR (2) remotely useful for assessing AI companies]. Our initial work will be on [concrete specific thing that would help].

METR has problem X. We should instead use preexisting organization Y, which does not have problem X and has comparable expertise at assessing loss-of-control risks and misalignment incidents inside AI companies, for instance [past work they've done of similar quality]. (this would be an amazing take but it is impossible to post because no such org exists)

OpenAI Suggests Doing The Least We Can Do

OpenAI has offered a new essay on the subject, Building standards for the next phase of AI. They emphasize that their goal is still an automated AI researcher, which then would be asked to do our alignment homework, but only if it can be done safely.

OpenAI: Fully autonomous RSI is not happening today, and we should not pursue it unless and until it can be done safely. Whether and how to proceed must depend on our ability to preserve human control and on informed democratic choices about the benefits and risks.

Done without appropriate care and caution, RSI could result in humans losing practical control over AI development, unable to provide oversight on research processes they no longer understand. From here, AI could become more dangerous, less aligned, and, on the whole, a danger to people.

We are at the point where OpenAI says they will ‘seek ways to keep humans in the loop’ of improvement, because otherwise that would not happen.

They next warn that fragmented international standards, or a lack of collective action, could lead to poor outcomes. Yes.

OpenAI: Our rationale for standards is rooted in avoiding the concentration of power, and producing better practical outcomes.

So many people feel the need to keep repeating ‘avoiding the concentration of power’ like it is a mantra, the concern that you are allowed to have. If we cannot get over that, I don’t see how we get to a good place.

OpenAI: We believe the United States should lead an effort to work together with countries around the world to develop global technical standards for frontier AI, including for RSI.

Okay, enough preliminaries. What is the proposal?

(1) A mechanism that facilitates complementary national and international frontier standards

They suggest leveraging the network of existing AISIs in many countries.

(2) Common measurements and incident reporting protocols for better collective action

Okay, sure, that’s all better than nothing, but this is a very milquetoast proposal.