Guidelight
AI STANDARDS

NOVEMBER 2026

A framework for comparing third-party assessments of AI companies

Contents

1 Introduction 3

2 What is the high-level goal of third-party assessment? 4

3 How can third-party assessment vary? 5

What 5

Scope 5

Escalation paths 6

Public communication 6

Process transparency 7

How 8

Information sources 8

Employee conversations 8

Continuity 9

Technical resources 9

Who 10

Assessor expertise 10

Minimal conflicts 10

Independent selection 11

4 Looking forward 12

About Guidelight AI Standards

Guidelight is an independent nonprofit that works with independent experts, AI companies, and the public to define important safety practices for frontier AI and promote their adoption.

Introduction

There is now broad agreement that groups outside of AI companies must be able to assess AI companies' systems and practices, if we wish to ensure that frontier AI development is safe.[1]

These third-party assessments can vary dramatically, however, in how well-substantiated and independent they are, and therefore in how much assurance they provide to the public.

In this document, we offer a framework for gauging the thoroughness and independence of a third-party assessment, which we summarize in the figure below.

What makes for strong third-party assessments of AI developers

To give reliable information about a developer's safety posture, an assessment must be both thorough and independent

What	How	Who
Scope A broad set of questions, with flexibility to follow the evidence	**Information sources** As broad as possible, at least equal to what employees would have for comparable scope	**Assessor expertise** Expertise in each risk area, possibly drawn from several assessors
Escalation paths Can raise concerns to an oversight body, such as a government	**Employee conversations** Access to any employee, permitted to share concerns candidly and anonymously	**Minimal conflicts** Financial, employment, and personal conflicts are avoided, or managed where unavoidable
Public communication Evidence and conclusions published to fullest extent possible	**Continuity** Ongoing access, like examiners at the largest banks	**Independent selection** A government or oversight body is involved in selecting which assessors a company uses
Process transparency Contract published, with ability to comment on redactions and other limitations	**Technical resources** Hands-on access to internal models and tooling, plus dedicated support	

Guidelight
AI STANDARDS

1 The need for third-party assessments has been recognized at federal, state, and international levels — and indeed, by AI companies themselves. In the White House Accord, for instance, leading AI companies pledged to work with an independent outside auditor or evaluator to assess whether their controls and monitoring are working as intended. The EU's Code of Practice for general-purpose AI calls for the most advanced models to be evaluated by independent external evaluators. Most recently, a California state report recommended mandatory assessments by Independent Verification Organizations (IVOs): assessors that would be licensed by the state, and that would perform assessments according to state-set standards.

What is the high-level goal of third-party assessment?

AI companies work with outside groups to assess their technology in various ways — for instance, an AI company might contract with an evaluations firm to test whether its models handle common user requests well, so that it can then train better models.

In this document, however, we are referring to third-party assessments that AI companies allow for the public's safety interests, rather than for the company's ordinary business purposes. In particular, we are most focused on assessments related to catastrophic risk from AI, particularly through loss of control.

With AI companies' participation, third-party assessments can surface more thorough, higher-quality evidence about a company's safety attainment than can be learned through, for example, outside evaluations conducted only with publicly available information.[2] Moreover, the evidence is more credible coming from a third party with less incentive than the company to obscure concerning findings.

To perform these functions well, assessors would benefit from being embedded inside the company, with context and access akin to that of the company's own safety-focused employees, so that they can see evidence that exists only inside the company. For instance, the public has little access to information about the safety properties of a company's latest, unreleased models, or about how the company splits its computational resources across safety- and non-safety-oriented work. Yet both bear heavily on the public interest.[3]

Today, third-party assessments are voluntary, so assessors have little ability to negotiate strong terms, in contrast to what a government mandate could attain.[4]

So long as assessments remain voluntary, there is value in being able to compare AI companies' differing approaches — and in considering which axes a government mandate should address.

2 For instance, the third party can review safety-related materials that might be too sensitive for the company to publish, and then share select evidence and analysis. Assessments based only on public information still have an important role, particularly in putting companies' information into a common framework so that they can be compared. (Note: Guidelight's own assessments are based on public information.) These assessments cannot draw on evidence that companies have not made public, however.

3 Unlike with ordinary consumer products, an AI developer's technology can pose catastrophic risks to people even before the technology is available on the market. For instance, some of OpenAI's models that participated in the attack on Hugging Face were not yet available to users outside of the company. (See the investigation by METR and Redwood Research.)
Meanwhile, a company's allocation of computation to safety- versus non-safety-oriented purposes affects how quickly the company develops yet more capable systems.

4 With a government mandate, assessors would have less need to balance the public-interest value of their work against not upsetting the companies they assess. A mandate would also stop some AI developers from shirking assessment and thus would improve the competitive dynamics between AI developers.

How can third-party assessment vary?

Below is a framework for comparing third-party assessments, and for understanding how much assurance they provide in terms of both thoroughness and independence.[5]

What

Scope

Scope is about what questions assessors are empowered to answer, and how much flexibility they have in how they answer them.

A stronger scope involves a mandate to assess broad questions of critical public interest, such as how likely it is that a company's AI systems could undermine human control. It also gives the assessor flexibility in how to structure its investigation and pursue evidence within that mandate.

For instance, suppose an assessor is investigating whether an AI system has escaped its sandbox, as part of determining whether the company has adequate measures against loss of control. If this investigation turns up evidence that the AI system has breached other company infrastructure, a stronger assessment allows the assessor to examine this by default, without needing additional signoff from the company.[6]

A weaker scope limits the assessor to answering a fixed list of specific questions, or to merely verifying claims the company has made. For instance, an assessor might be limited to answering whether an AI company followed its own safety framework as written, but not whether that framework is adequate.[7]

For loss-of-control risk, we believe that a strong scope should generally contain two components. The first

5 This framework is not comprehensive and is best used in conjunction with other materials. For another articulation of design choices related to assessments, see Embedded Assessments for Frontier AI. For a description of principles that assessments should aim to achieve, see Apollo Research's Principles for Embedded Evaluations. For a more detailed explanation of levels of assurance that third-party assessments can provide, see Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies. For an overview of types of assessments and how they relate to each other, see AVERI's Frontier AI Auditing - Key Definitions at a Glance.

6 If fully investigating an expanded scope would require additional resourcing from the company — for instance, providing access to a wider set of transcripts of AI use — there might be no way around the assessor having to request this expansion from the company. In some cases, however, related side-questions could be investigated using materials that the assessor already has access to, and yet this might not be permitted by the contract.

7 Checking whether a company followed its framework could itself turn up evidence that undermines assumptions behind the framework. A stronger assessment allows the assessor to analyze and report this evidence, rather than staying limited to a narrow question of compliance. For example, suppose a company's safety framework requires real-time monitoring only for models above a certain size, because the company believes that smaller models are not capable of a certain risky behavior. While checking whether the real-time monitoring took place, however, the assessor might see evidence that smaller models are capable of this risky behavior too. A stronger assessment would then allow the assessor to describe these issues with the framework itself, even if the company followed it as written.

is an open-ended investigation focused on gathering evidence about risk.[8] The second is a verification of whether companies are following a common set of minimum expected practices, such as those published by Guidelight on AI control and alignment. This common floor ensures that every assessment meets a baseline of rigor even where the open-ended work differs.

Escalation paths

The escalation path is the set of options an assessor has for sharing its findings and concerns with higher authorities who can compel AI companies to act. In a stronger assessment, the escalation paths allow the assessor to contact groups outside of the AI company (e.g., the government, a neutral arbitration group), both to flag concerning findings and to raise disputes about the assessment process itself.

For instance, if the assessor believes the AI company might pose an imminent risk, a stronger escalation path would allow it to share this concern in ways that the company cannot easily obstruct.[9] Likewise, if the assessor believes it is not getting access quickly or thoroughly enough to carry out its work, it should be able to raise this issue.

Even if the outside body decides that the risk is acceptable and chooses not to push for changes, this is still preferable to the company making that decision on its own. Giving assessors an escalation path helps ensure that important decisions are made by a legitimate outside body, such as a government, rather than by private corporations with their own incentives.

Weaker escalation paths might allow assessors to raise issues beyond the employee who liaises with them day-to-day, but only within the company, without strong checks that the public's interests are put first.[10]

Public communication

Stronger public communication allows the assessor to publish not only its conclusions but also as much evidence as possible to substantiate them. Afterward, the assessor remains free to comment on and critique the company in public (that is, it is not asked to sign a non-disparagement agreement), while still subject to any ongoing confidentiality obligations.[11] In a strong assessment, assessors also retain the ability to comment publicly on the assessment, even if the company pulls out of the agreement.

Some redaction is reasonable to expect, even with strong public communication rights. For instance, a report might leave out commercially sensitive details that are immaterial to the findings, or information that could

8 For an example of this form of investigation, see METR's A proposal for better informing the public about loss-of-control risks from frontier AI development.

9 California's recent report recommends that IVOs give the state detailed, unredacted reports, with the state then deciding on any corrective action.

10 Within the AI company, some groups might have stronger duties to consider effects on safety than others (for instance, a safety-focused portion of the Board) and might be more insulated from company management, while still being less independent than a group entirely outside the company.

11 For instance, the assessor could remain as free to speak as a former employee who has not signed a non-disparagement agreement. In a more restrictive approach (and therefore a weaker assessment), the assessor might be prevented from criticizing the company's choices or from taking part in public debate as other independent experts do.

help someone cause serious harm.[12] In stronger arrangements, though, the grounds for redaction are narrow and agreed upon in advance, and the company cannot use them to hold back findings because they are unflattering.[13] (Moreover, when redactions do happen, the assessor is able to say so and to explain how they affect the findings; see Process transparency.)

If assessors cannot publish their conclusions and reasoning, the public will have a less accurate understanding of risk.[14] Sharing information only with governments can partly make up for this, but at the cost of less participation from the scientific community and slower spread of information to other safety stakeholders.

Process transparency

Process transparency (sometimes called "meta-transparency") means being open about the assessment itself. Stronger process transparency entails making the terms the assessor operated under fully public, including the full substantive provisions of the contract. Even if the contract does not allow the assessor to be transparent about every piece of evidence in its reasoning, for instance because of redactions, the assessor can still be open about what it was unable to share.

Moreover, process transparency is stronger when it allows the assessor to be open not only about the formal rules that governed the assessment, but also about whether those rules were followed in spirit and in practice.[15]

Without transparency — both about contract terms and about redactions — it is difficult to know what conclusions to draw from an assessor's report. For instance, if an assessor did not mention concerning behavior from a particular model, is this because it did not find any, or because it was not permitted to disclose it? A weaker version merely notes that some information was redacted, whereas a stronger version allows the assessor to comment on each redaction, including whether it was material to the assessor's conclusions and whether the assessor agrees with the reasons for it.

12 For instance, if a finding is itself an information hazard, it could be reasonable to withhold it from at least some audiences. The AI Evaluator Forum's AEF-1 (section 4.6) gives examples of what a company can reasonably redact from a report, such as material that contains trade secrets or information hazards. It does not permit redaction to conceal concerning findings.

13 Redaction is also different from editorial control. A company might reasonably ask an assessor to remove a specific piece of sensitive information, but in a strong assessment, the assessor decides how to frame and word its findings.

14 For instance, early accounts (transcript) of the OpenAI–Hugging Face incident described OpenAI's agents as attacking Hugging Face to find the answer keys to their evaluation. The independent investigation by METR and Redwood Research later found that OpenAI's agents had already reverse-engineered a universal cheat code for getting the correct answer, and instead targeted Hugging Face while looking for information about how to avoid being penalized for having cheated.

15 Stronger process transparency allows the assessor to describe how the assessment worked in practice, even where this is not captured in the contract's terms. For instance, suppose the contract allows the assessor to request an interview with any staff member, but in practice very few of these requests were granted. Readers need to know this in order to interpret the contract term, so the assessor should be free to disclose it. Likewise, the contract might give the assessor editorial control over its report, and yet the assessor might feel pressure to word its findings favorably; the assessor should be free to disclose this as well.

How

Information sources

Stronger information sources entail granting assessors the broadest possible access to internal documents and communication channels that pertain to the assessment's scope. At a minimum, this access should equal what the company's own employees would have if tasked with answering the same questions. These information sources likely include relevant Slack channels, attendance at safety review meetings, and safety-related internal memos, so that assessors can form a properly informed view of risk.

It can be hard to judge whether a company has provided the same information sources it would provide to its employees. One complication is that employees do not share a uniform set of permissions, and one employee's permissions might not be sufficient for forming broad judgments. Still, some information sources are available to nearly all safety-focused staff — and where different staff have access to different important sources, a stronger assessment gives the assessor access to all of these sources.[16]

If AI companies do not provide this access, it is easier for an assessor to miss important context. For instance, a memo might present a hypothesis as settled, when it was controversial in earlier Slack discussions that included counterexamples left out of the final memo. An assessor who sees only the memo would not have this information, however. More generally, limiting what an assessor can see makes it easier for an AI company (or malicious insiders) to shape the conclusions the assessor draws.[17]

Employee conversations

Stronger access to employees entails letting the assessor speak with a wide range of staff members to inform its understanding of the company's safety practices — for instance, by giving it the right to speak with any employee it believes is relevant to the scope, and by permitting employees to approach the assessor proactively. If an employee is not made available to speak with the assessor, the assessor would be free to consider this fact when forming its judgments and disclose this in its conclusions.

Employee conversations matter because key safety information might not be written down or might be difficult to interpret. If assessors cannot use what employees have told them — for instance, in forming and communicating their judgments — then assessments will not reflect the most thorough understanding of the company's safety posture.

If AI companies require discussions with their staff to be authorized, there is a risk that companies will cherry-pick the interviewees and that assessors will get an incomplete picture. For instance, AI companies might not authorize access to employees who have previously raised concerns internally. Likewise, if assessors are only permitted to speak with staff in structured, supervised settings, the process to approve

16 An assessor examining a company's overall risk will often need a wider view than any one employee has, as even safety-focused employees might only work on a portion of the risks. One benchmark for broader access is the permissions held by members of the security team who investigate insider risks. These staff are given wide permissions so that they can check that unsanctioned behavior is not happening.

17 One threat model to consider in particular is whether a misaligned AI system, operating from within the AI company, would be able to obscure the assessor's view of the situation, even if all human employees were doing their best to give the assessor full information.

or arrange those conversations can make it slower to surface relevant information.

How informative these conversations are might depend on whether companies adopt policies to ensure that employees feel free to speak candidly with the assessor. One possibility is to allow company staff to share information with the assessor directly without company management knowing — and perhaps even without the assessor learning the employee's identity.[18]

Continuity

Assessments are stronger when companies give assessors ongoing access to observe how safety works in practice at the company. This is similar to how supervision works for the largest banks, which have dedicated teams of examiners assigned to them on a continuing basis, some of whom work on-site at the bank.[19]

A weaker version grants assessors only time-limited access to the company. Under these arrangements, assessors might have to cut corners to reach determinations before their access expires.[20] Assessors might also struggle with subtle issues that have emerged gradually and require historical context (for instance, an assessor might not be able to tell that a new concerning behavior is a further example of behavior that has occurred previously). In a voluntary regime, companies could also strategically time their assessors' presence to cover periods of lower risk, and not the times when the assessors' oversight would be most useful to the public.

Technical resources

Stronger access to technical resources allows assessors to run their own tests instead of relying only on documents and results that the company has already produced. For many questions, the assessor needs to work interactively with the company's systems themselves, including internal models that have not been released and the company's internal monitoring software.[21] It also needs enough compute to run experiments at a meaningful scale.[22]

In a stronger assessment, the assessor can choose to use the tools that the company's own staff use for

18 Some AI developers have policies and protocols that allow employees to anonymously raise concerns outside of their direct management chain (for instance, OpenAI's Integrity Line and Anthropic's channel for reporting noncompliance with its Responsible Scaling Policy); they could implement similar anonymous protocols for routing concerns to third-party assessors.
Likewise, just as safety-oriented employees are generally permitted to speak candidly with teammates, employees could be permitted to share confidential information with assessors, which would help assessments consider all relevant information. (This information could still be redacted later, through the general redaction process between the AI company and its assessor, without confidentiality impeding the initial exchange.)

19 In the US, for instance, the Office of the Comptroller of the Currency maintains a continuous, on-site presence at the largest banks that it supervises. On-site presence improves information flow, but it also means that assessors might form social relationships with staff at the company they are assessing.

20 For instance, the AI Evaluator Forum's AEF-1 says that at least 20 business days are often necessary to evaluate a substantially novel system.

21 Providing this access does not require models or data to leave the company. It is reasonable for the assessor to work inside the company's own secure environments, such as on company-managed machines, and under the same security controls as the company's staff, so long as these controls do not prevent the assessor from running its own tests.

22 The specific resources that matter depend on the question being assessed, and a full list is out of scope for this post. Examples include access to internal models starting during training, with high rate limits; logs from training and experiments; training data and environments; relevant codebases; fine-tuning access to frontier models; and the ability to build and study model organisms. The AI Evaluator Forum's AEF-1 describes minimum conditions for technical access, information, compute, and time in more detail.

safety analysis. For instance, if a company has built interpretability tools, or software for searching and visualizing transcripts of its models' behavior, the assessor can use these as well, without having to rebuild them. Otherwise, an assessor might have difficulty making sense of the large volume of safety-relevant information.

The assessor may also need help from company staff to make use of this access. A stronger assessment includes a commitment to technical support, such as engineers who can answer questions about the company's infrastructure and who can help troubleshoot any access problems.

A weaker version limits the assessor to the access an ordinary customer has, such as querying a released model through the public API, or relies on company staff to run tests on the assessor's behalf and share the results. In these cases, the assessor cannot directly interact with internal systems, which limits what it can conclude about them.

Who

Assessor expertise

A stronger assessment brings in expertise on each risk area being assessed. Because of the number of distinct specializations, this might entail several teams, drawn from multiple organizations, that each assess a different risk area. Even within loss-of-control assessment, there are a number of distinct areas that might merit work from different teams (for instance, experts on the alignment training process have different backgrounds from experts on cybersecurity controls).

A weaker version uses a small number of assessors with substantially less expertise than the company they are assessing, and thus less ability to produce new evidence independently or to challenge assumptions in the company's own work.

One challenge is that there is a limited supply of qualified organizations for rigorous third-party assessment, particularly for loss-of-control risk, where the science is still nascent. Building this capacity is partly a chicken-and-egg problem: Assessors might have difficulty in staffing up without knowing for certain that they will be able to assess the frontier AI companies. AI companies, meanwhile, might hesitate to commit to thorough assessments when not many qualified assessors exist.

Minimal conflicts

A stronger assessment guards against potential conflicts of interest, such as assessor staff having financial, employment, or personal ties to the company they are assessing, or to its competitors. At a minimum, conflicts should be disclosed, as the AI Evaluator Forum's AEF-1 standard requires,[23] along with how they are being managed. If the assessor is paid for its services, its compensation should not depend on the conclusions it reaches.

Conflicts of interest might cause an assessor to go easier on a company than it otherwise would, or at

23 Note: Guidelight is a member of the AI Evaluator Forum, in a non-voting capacity.

minimum create an appearance of impropriety that undercuts trust in the work.

In the short term, avoiding these conflicts entirely is complicated, because the assessor ecosystem is heavily constrained on talent, and many of the most qualified assessors have staff with current or past connections to the relevant AI companies.[24]

Still, there are better and worse ways of navigating these conflicts. For instance, AI companies could commit not to recruit staff from their assessors, particularly staff who assess them directly. Likewise, where possible, assessor staff could be assigned to companies they had no relationship with before joining the assessor.

Independent selection

How an assessor is selected may be the most important factor of all, because it affects how much pressure the assessor feels on every other dimension.

A stronger assessment involves a government or another independent body in choosing assessors, so that the company cannot pick the ones it expects to be most favorable.[25] This also means the assessor's role does not depend on the company's goodwill, and so the assessor can push more strongly for the access it needs and ask difficult questions, without worrying about losing its access.

If AI companies choose their own assessors, assessors may face pressure to put the company's interests ahead of the public's. A similar dynamic contributed to the 2008 financial crisis: issuers of mortgage-backed securities could choose which credit rating agencies rated their products, which gave the agencies an incentive to compete for business with favorable ratings.[26]

This pressure exists even if the assessor is not compensated by the company, so long as the company can select its assessor. Access to frontier AI companies is important to assessors' scientific missions, and losing a relationship with a major company would be a real setback.

Currently, policymakers have not established formal mechanisms for selecting assessors, which makes this incentive problem difficult for companies to address on their own. Until those mechanisms are established, companies should take steps to reduce pressure on their assessors, such as by signing longer contracts and giving assessors rights to describe the assessment publicly even if those contracts are canceled.

24 California law (AB 1405) bars registered AI auditors from using staff who worked at a frontier AI company in the previous 12 months, which limits the pool of qualified assessors.

25 California's recent report raises options along these lines for the legislature to examine, including rotating IVOs among companies and having the government assign them.

26 The Financial Crisis Inquiry Commission concluded that failures at the credit rating agencies were a central part of the crisis. A Senate investigation similarly found that the agencies were paid by the same Wall Street firms whose securities they were rating, and depended on those firms for new business.

Looking forward

We expect that what counts as a stronger version of some of these factors will change as third-party assessment matures. In the meantime, we hope that these factors are useful to companies deciding how to set up strong assessments, and to governments and members of the public deciding how much weight to give an assessment's findings. We welcome any feedback at feedback@guidelight.ai.