METR

Conflict of interest policy (version 1.0)

Last updated August 28, 2026

METR is a nonprofit which conducts research and evaluations to improve public understanding of the capabilities and risks of frontier AI systems.

This policy concerns projects in which decision makers or the general public may rely on our work with a frontier AI company[1] to understand what level of risk that specific company is posing. Here, we'll refer to projects used this way as "company-identifying risk assessments", as distinct from research on risks broadly.[2] As the possibility of loss of control becomes more imminent, the accuracy and impartiality of such risk assessments will become increasingly important.

In this policy, we commit to processes for managing conflicts of interest which may impact the reliability of our conclusions. Primarily, the policy covers:

- The key test and a process for identifying and determining responses to any conflicts which may arise in a company-identifying risk assessment.
- A commitment to disclosing unmitigated conflicts in that project's external-facing report(s), and what mitigations, if any, are sufficient for different kinds of potential conflicts.
- Who specifically is responsible for carrying out these actions (named in **bold**).

Policies regarding individual staff

How to understand the materiality of a conflict of interest

Write down the full details of any potential conflict(s) of staff working on the risk assessment, and how those staff contributed. If we published that in full, would a reasonable person not affiliated with METR have materially less trust in the assessment?

---

[1] By frontier AI company, we mean the AI companies developing models with the strongest autonomous capabilities (for which METR might run a high-stakes risk assessment, or do so with a direct competitor). Examples as of July 2026 include Anthropic, Google DeepMind, Meta Superintelligence Labs, OpenAI and SpaceXAI, but these may change over time. Please consult your manager if you are unsure whether an entity is in scope.

[2] Some notable examples of similar work which we do *not* class as company-identifying risk assessments include:

- Developing risk models and risk assessment methodologies.
- Work on infrastructure used in risk assessments.
- Combining information about several companies into a report about risk across the industry which does not discuss the level of risk posed by specific companies.

If in doubt about how to manage a potential issue, return to this test.

Main process

When setting up a company-identifying risk assessment, **the individual responsible for determining who works on the company-identifying risk assessment** will:

1. Hand this responsibility to another individual if you have a Tier 2 or higher conflict as specified below.
2. Require potential project staff to:
    - Share any interest of theirs which might cause a reasonable person not affiliated with METR to have less trust in the project's conclusions, using this form.
    - Report potential policy violations.

    Optionally, you can adapt this template[3].
3. If any potential project staff has a conflict, determine what team selection and mitigations would maximize the reliability of the risk assessment for its intended audience, balancing required skills and potential bias. (Review the example conflicts and disclosure requirements to calibrate your approach. Also, check that the proposed team is compliant with applicable external rules.) If maximizing the reliability of the risk assessment for its intended audience may require proceeding with conflicts which require external disclosure, prepare external disclosure.
4. Write simple documentation[4] of the decisions made, in case needed for future review. (See review of potential violations.) Share this with the most senior other unconflicted individual read in to the company-identifying risk assessment.

If you would benefit from assistance, consult METR's operations lead, who oversees METR-wide implementation of this policy and can also direct you to others who could help.

Detail: Example conflicts and external disclosure requirements

Scope: Disclosure requirements apply to staff who make or oversee a material contribution to subjective judgements about the company which would reasonably be expected to materially impact the conclusions of the risk assessment. Examples of functions which may be sufficiently objective and thus out of scope include:

- Building infrastructure and tooling.
- Executing predefined benchmarks or other predefined data generation.

[3] *Links to internal forms have been removed in the public version.*

[4] For example, this could simply be a statement that the project team will follow the rules laid out here in full, any actions that requires, and links to the relevant staff's disclosures. If there is a conflict and you decide to apply weaker mitigations than standard, you should additionally write down this decision, including any restrictions on how the individual is allowed to engage with the project, how well you expect this to mitigate risk of bias in the project's conclusions, and your evidence and reasoning for this.

The following table describes examples of conflicts which could arise between METR staff and frontier AI companies, and under what circumstances they must be disclosed in that project's external-facing report(s).

Example conflicts which could arise between project staff and a company subject to a company-identifying risk assessment	**Circumstances under which such a conflict requires <u>external</u> disclosure** – in that project's external-facing report(s)
Tier 3 ● Current employment, contracting or other paid work with the company, or pursuit of such work (including through informal conversation[5]). ● Ongoing agreements which may impair frankness, such as non-disparagement. ● Current direct equity or debt holdings[6] in the company or a reasonable economic proxy for that specific company,[7] or any other financial instrument of similar effect,[8] including interests via marriage. (Exposure via indices is not prohibited.) ● A current romantic relationship with an employee of the company. ● Immediate family employed at the company. ● Current indebtedness to the company or one of its employees.[9] ● Gifts from an employee of the company which sum to greater than $100 in value from that individual in the 6 months	Always requires external disclosure. Should you decide to involve a person with a Tier 3 conflict, consider whether the Tier 2 mitigations, below, would be appropriate.

[5] That is, during a risk assessment project with that specific company, you should always decline to discuss employment even if asked, else this would require public disclosure and you may be required to drop out of the project. Example response if asked about a role while you are working on, or may soon work on, a project for which this would pose a conflict: "To minimize conflicts in the kind of work I do, I'm currently not considering roles with frontier AI companies."
[6] Examples: vested or unvested equity, stock options, deferred compensation interests, or security interests such as SAFEs.
[7] That is, assets for which you see a large part of their value being their exposure to a specific AI company. Examples: Softbank/Oracle or SK Telecom/Zoom.
[8] This includes digital assets and prediction markets.
[9] This excludes amounts owed in the ordinary course of splitting shared expenses.

preceding the company-identifying risk assessment.[10]	
Tier 2 ● A Tier 3-type conflict with a direct *competitor* of the company. ● Currently living with an employee of the company. ● A historical romantic relationship with an employee of the company which ended in the last 12 months.	Requires external disclosure unless an unconflicted staff member thoroughly[11] checks[12] and takes personal responsibility for the accuracy of all work by this individual on or incorporated into the main report.
Tier 1 ● Close friendship with an employee of the company.	Requires external disclosure unless there is no contact with said friend about the project (even if the friend is on a relevant team).

A conflict might arise which does not clearly fit into any of these categories. If in doubt about how to handle a potential issue, **return to the key test**.

Detail: Preparing external disclosure

For risk assessments requiring extremely rare expertise, proceeding with a conflict requiring disclosure may be worthwhile in order to bring that expertise to bear.[13] However, the audience of the risk assessment (often including the general public) must be informed of such weaknesses through disclosure in that project's external-facing report(s).

If you choose to proceed with a conflict requiring disclosure:

[10] This excludes shared expenses, such as a shared meal or other shared item, so long as you later reimburse the other party for your part.
[11] Thorough checking should follow at least the usual quality of review of METR research, entailing understanding the data and analysis well enough to be able to understand the validity of the conclusions, and including consistency checks, understanding the provenance of statistics, and reading through some of the underlying data.
[12] The individual responsible for determining who works on the company-identifying risk assessment should make a note to use the following process or another in good time before the end of the project to ensure any such content is checked: (1) Assign an unconflicted individual to tag components of the final report or work incorporated into it to which the conflicted staff member contributed, using interviewing contributors, Git tools and Google Docs history as needed. (2) Assign unconflicted individual(s) to thoroughly check and take responsibility for these sections (commenting on the main report that they have done so).
[13] In addition, there may be unforeseen cases in which our disclosure requirements are overly conservative. One heuristic for this is whether we would see negligible risk of bias even if the role in question was carried out by staff of the company itself. We should nonetheless disclose such cases.

1. Write down language for disclosing the conflict(s) in that project's external-facing report(s). The individual's identity should be anonymized to the extent practical. Make a note to ensure that this language is included in those report(s).
    - This is a key tool for ensuring the benefits of proceeding with an unmitigated conflict truly outweighs its costs.
    - If there are multiple disclosures between project staff, consider the disclosures in their totality. (See the key test.)
2. Require the most senior other unconflicted individual read in to the company-identifying risk assessment to review and approve the documentation of the decision and the disclosure of unmitigated conflicts. They should consider:
    - Whether the proposed disclosure provides enough information for a reasonable outside party to understand approximately how much risk of bias is incurred.
    - Whether the participation of the affected individual will overall make the project more trustworthy, when considering both the expertise they bring to the project and the reality or perception of reduced impartiality.
3. Provide the disclosure to our point of contact at the company that the project is focused on (if the project is a collaboration).

Review of potential violations

When **METR's operations lead**, the **individual responsible for determining who works on the company-identifying risk assessment** or another staff member receives a notification of a potential policy violation, they will report those potential violations to the members of the leadership team who are read in to the company-identifying risk assessment.

The members of the leadership team who are read in to the company-identifying risk assessment will review any potential violation of this policy, unless it potentially impacts the CEO's impartiality, in which case it will be reviewed by independent members of the board. This review will:

- Judge whether the individual should be removed from the company-identifying risk assessment or should have limits set on how they may engage with the project.
    - Provisional recusal will be used prior to the review unless the concern is low-confidence.
    - In the case that there are no appropriate reviewers read in to the project, the review will occur after the project is complete, and therefore full recusal from the project will therefore be necessary.
- Determine whether a deliberate or reckless violation occurred which warrants disciplinary action, up to and including termination.
- Determine what process changes should be made to prevent future such violations.

Ensuring individual staff understand the policy

METR's operations lead will require any staff who may be interested in contributing to company-identifying risk assessments, upon onboarding and semiannually, to:

- Review the policies regarding individual staff, and document any potential conflicts to ensure they understand the policy's potential implications for their work.
- Either:
  - Share the documentation with their manager and [address][14], or
  - Confirm to their manager and [address] that they have documented any potential conflicts privately and understood their potential future implications, but decline to share that documentation.[15]
- Report potential policy violations via a standard form.

Additionally, remind staff of which of their colleagues they may wish to speak with if they have questions.

Practices regarding METR as a whole

METR's CEO will continue to promote a culture of impartiality. Cultural matters are necessarily less clear-cut than the personnel policies outlined above. However, a notable example of concrete implementation is METR's core principles:

- Core principles which guide METR's strategy are typically reviewed in team strategy days approximately quarterly. They feed into the selection and execution of METR projects.
- One such core principle is *"truthseeking, trustworthy"*. That is, METR attempts to figure out the truth and communicate it accurately.
  - Examples of the care METR puts into communicating findings as accurately as possible can be found in Notes on Scientific Communication at METR.
- Another core principle is *"honest about important things, even if weird"*. This builds on the *"truthseeking, trustworthy"* principle, emphasizing that METR does not censor its communications when it discovers surprising evidence.
- Another is *"not zero-sum / advancing one group over another"*. That is, METR neutrally assesses risks, and never sets out to impact the relative position of companies.

METR's finance lead will ensure that METR itself continues to not be financially dependent on frontier AI companies:

---

[14] As of August 2026, [address] is accessible to METR's head of operations. It will be accessible to other staff responsible for METR-wide risk management in the future.

[15] Sharing this documentation enables METR to plan ahead for staffing important projects, and prevents surprises at the point of launching such projects. However, to protect employee privacy, we only *mandate* sharing when you are under consideration for joining a company-identifying risk assessment.

- Do not invest in direct equity or debt of frontier AI companies or reasonable economic proxies for specific companies.[16]
- Decline donations made by or at the direction of[17] frontier AI companies or their employees.[18,19] Use of free compute credits is acceptable.
- To date, we have not received payment for work on company-identifying risk assessments (or other work with frontier AI companies). If we ever do, the payment terms must not be contingent on the results of the risk assessment.
- (More generally, METR strives to be supported by broad and independent funders.)

METR's operations lead will ensure METR is not directly answerable to frontier AI companies. Concretely:

- Employees of frontier AI companies, or individuals with a substantial fraction of their net worth invested in a frontier AI company, are not eligible to serve on METR's board.

Revisions

We may revise this policy to better balance staffing projects with highly capable staff, protecting our impartiality, and executing efficiently. If so, we will publish the updated policy by the time it goes into effect.

---

[16] Details as per the rules for employees.
[17] This prohibits both direct donations and situations like a frontier AI company employee who is a board member of or advisor to a public charity (e.g. DAF provider) instructing that charity to donate to METR. It does not prohibit funding from a public grantmaking organization that itself has received funding from frontier AI company employee(s), so long as such employee(s) are uninvolved in grantmaking decisions.
[18] For example, we have declined grants offered by the founders of multiple AI companies.
[19] In the rare case that it is unclear who directed a donation, we will seek confirmation that it was not directed by a proscribed entity or individual if it amounts to 0.1% or more of METR's annual income.