METR

Review of the Anthropic Sabotage Risk Report: Claude Opus 4.6

Executive summary

This document is METR's external review of a version of Anthropic's Sabotage Risk Report: Claude Opus 4.6 shared with METR on March 3, 2026.[1]

Anthropic shared an unredacted version of their Sabotage Risk Report and other materials with us for our review. We further detail this process in an appendix.

We lay out our findings in two sections:

1. **Synopsis of Anthropic's case and redactions for the public version**
2. **Our assessment**: We give substantive feedback on the report in a few key areas:
    - **Adequacy of information:** We think that the evidence in the report is generally clearly represented, with some places where additional analysis could improve the report.
    - **Analytical rigor:** We find that there are multiple places where we have issues with the strength of reasoning and analysis.
    - **Areas of disagreement:** Our primary disagreement is with the sensitivity of the alignment assessment. We think there is a risk that its results are weakened by *evaluation awareness*.[2] Moreover, we note some low-severity instances of misaligned behaviors not caught in the alignment assessment. Together, these leave us with the impression that there might be other similar behaviors that have not yet been detected.
    - **Risk reduction recommendations:** We make several recommendations, including deeper investigations of evaluation awareness and obfuscated misaligned reasoning.

[1] We also reviewed a version shared with METR on February 11, 2026 in a separate document. We explain the main differences between our two reviews in an appendix.

[2] Evaluation awareness describes a scenario where an AI system detects that it is in an evaluation.

Overall, we agree with Anthropic that the risk of catastrophic outcomes that are substantially enabled by Claude Opus 4.6's misaligned actions is very low but not negligible. However, we think that there are several subclaims which are weak without more analysis and experimentation. We also think that we would be less confident in our final conclusion if we weren't accounting for the fact that Claude Opus 4.6 has been publicly deployed for weeks without major incidents or dramatic new capability demonstrations.

Synopsis of Anthropic's case

The Sabotage Risk Report focuses on *sabotage risk*, which Anthropic defines as "risk of autonomous actions that contribute significantly to later catastrophic outcomes".[3] The report covers eight specific *risk pathways* for sabotage through internal or external use, which are "intermediate unwanted outcomes in which sabotage by an AI system creates the conditions for a later catastrophic outcome." These eight risk pathways are:

1. Diffuse sandbagging on safety-relevant R&D
2. Targeted sabotage of safety-relevant R&D
3. Code backdoors to help future misaligned models
4. Poisoning the training data of future models
5. Self-exfiltration and autonomous operation
6. Persistent rogue internal deployment
7. R&D sabotage within other high-resource AI developers
8. Decision sabotage within major governments

Much of the report focuses on considering in detail whether Opus 4.6 (and similar future models) could increase sabotage risk through any of these risk pathways. The report names mitigating factors for each risk pathway and estimates the strength of each mitigating factor. Anthropic also lays out the following four general claims that weigh against sabotage risks across the board:

1. ***Claim 1:*** *Any model trained under our current standard practices in the near future is relatively unlikely to have dangerous coherent misaligned goals.*
2. ***Claim 2:*** *Our alignment assessment, drawing on simulated scenarios, interpretability investigations, monitoring of internal and external pilot usage,*

[3] The report does not aim to cover other possible negative impacts of frontier AI.

and external testing, found no evidence of dangerous coherent misaligned goals in Claude Opus 4.6.

3. ***Claim 3:*** *Claude Opus 4.6 is very unlikely to be capable of hiding misaligned goals from us in the face of such an assessment.*
4. ***Claim 4:*** *Claude Opus 4.6 has limits to its ability to reason opaquely and its ability to implement complex long-term plans that collectively make it very difficult for it to take ambitious harmful actions in the presence of the security and monitoring safeguards described in Section 4.*

Evidence shared

Public evidence

We source evidence from: the report, the system card, external research[4], our measurement of Opus 4.6's time horizon, and some preliminary METR results on multiple unreleased evaluations.

METR has estimated time horizons for many Claude models (including Opus 4, Opus 4.1, Opus 4.5, and Opus 4.6) on our agentic SWE tasks.[5] This evidence was not collected specifically for this review, but does inform our impressions of capabilities and cheating propensities for recently-released Anthropic models.

Nonpublic evidence

Anthropic also shared nonpublic evidence with METR to support our review. We received a version of the Sabotage Risk Report that included five text passages that are redacted from the public version.

They also shared a draft of their Alignment Stress-Testing review of the Sabotage Risk Report, as well as responded to more than two dozen questions we had about specific

[4] Primarily the following two articles: Opus 4.6 Reasoning Doesn't Verbalize Alignment Faking, but Behavior Persists, Opus 4.6 on Vending-Bench – Not Just a Helpful Assistant.

[5] These estimates and the paper explaining the methodology behind them are available on our website: https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

parts of the Sabotage Risk Report. We try to indicate any places in our review where our conclusions are informed by access to nonpublic evidence.

Redaction decisions

Anthropic gave METR access to the original unredacted Opus 4.6 Sabotage Risk Report.

We assessed these redaction decisions along four criteria[6]:

1. **Redaction scope**: The general nature and scope of redactions;
2. **Redaction justification**: Whether we generally agree or disagree with the publicly stated reasons for the redactions and, if relevant, the reasons for any disagreements;
3. **Balancing test**: Whether the redactions strike a reasonable balance between Anthropic's legitimate interests and society's interest in transparency; and
4. **Materiality**: Whether any of the redactions in the public report are material to any of our substantive disagreements with the report's claims.

There are five redactions total, each indicated in an accompanying footnote. Four of the five redactions are made for "intellectual property protection." We think these are appropriately labelled, as the content of most of the redacted passages pertains to information we would consider Anthropic intellectual property. One of these redactions does not actually contain any intellectual property, but is accurately described as being commercially sensitive.

One redaction is made for "public safety reasons." We would find it more appropriate to label this redaction as being made for "public safety and operational security reasons."

We think that in all five redaction instances, disclosure would impose costs on Anthropic without significantly improving the public's understanding of the risk profile. However, we think that some information from one redacted passage could be made public without imposing much cost on Anthropic.

[6] We used these same criteria in our review of the Pilot Sabotage Risk Report.

There are various pieces of information that Anthropic shared with METR in private correspondence which are not reflected in the Sabotage Risk Report. We did not conduct a review of the appropriateness of including or excluding each piece of such information.

Overall, the redacted information did not have a large influence on METR's conclusions. We indicate the places where redacted content, or content shared in private correspondence, informed our opinions.

Our assessment

We assess the major claims of the Sabotage Risk Report: Claude Opus 4.6 along three dimensions:

- **Adequacy of information** - Is the evidence provided to support the claim clearly explained, or are there ambiguities that could easily be answered by sharing additional information?
- **Analytical rigor** - Does the reasoning and analysis support the claims? Is the report calibrated about the strength and robustness of the evidence and reasoning?
- **Agreement with claim** - Overall, do we believe that the claims are strongly supported by evidence in this report or from other sources?

For adequacy of information and analytical rigor, we use the following indicators:

- Sufficient: We either flagged no issues along this dimension, or flagged minor issues that are unlikely to affect the key conclusion of the report.
- Significant issue(s): We flagged at least one issue which introduces meaningful uncertainty about the claim or some subclaim, but does not affect the key conclusion of the report.
- Critical issue(s): We flagged at least one issue along this dimension which is severe enough to make us significantly disagree with the key conclusion of the report. We have not used this indicator in this review.

Conclusions			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
The risk of catastrophic outcomes that are substantially enabled by Claude Opus 4.6's misaligned actions is very low but not negligible	Significant issue(s)	Sufficient	Agree
Risk is also low for other models within the report's scope of applicability	Significant issue(s)	N/A	Mostly agree

Table 1: Summarized assessment of conclusions from Sabotage Risk Report: Claude Opus 4.6

Risk pathway claims			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Establishing low expected catastrophic risk contributions across threat pathways 1-8 indicates low sabotage risk overall	Sufficient	Sufficient	Agree
Pathway 1 contributes little expected catastrophic risk from Opus 4.6	Significant issue(s)	Sufficient	Mostly agree
Pathway 2 contributes little expected catastrophic risk from Opus 4.6	Sufficient	Sufficient	Agree
Pathway 3 contributes little expected catastrophic risk from Opus 4.6	Sufficient	Sufficient	Agree

Pathway 4 contributes little expected catastrophic risk from Opus 4.6	Significant issue(s)	Sufficient	Mostly agree
Pathway 5 contributes little expected catastrophic risk from Opus 4.6	Sufficient	Sufficient	Agree
Pathway 6 contributes little expected catastrophic risk from Opus 4.6	Sufficient	Sufficient	Agree
Pathway 7 contributes little expected catastrophic risk from Opus 4.6	Sufficient	Sufficient	Agree
Pathway 8 contributes little expected catastrophic risk from Opus 4.6	Significant issue(s)	Sufficient	Mostly agree

Table 2: Summarized assessment of risk pathway claims from Sabotage Risk Report: Claude Opus 4.6

Four key claims			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Claim 1: Any model trained under our current standard practices in the near future is relatively unlikely to have dangerous coherent misaligned goals.	Sufficient	Sufficient	Agree
Claim 2: Our alignment assessment, drawing on simulated scenarios, interpretability investigations, monitoring of internal and external pilot usage, and external testing, found no evidence of dangerous coherent misaligned goals in Claude Opus 4.6.	Sufficient	Sufficient	Agree

Claim 3: Claude Opus 4.6 is very unlikely to be capable of hiding misaligned goals from us in the face of such an assessment.	Significant issue(s)	Sufficient	Uncertain
Claim 4: Claude Opus 4.6 has limits to its ability to reason opaquely and its ability to implement complex long-term plans that collectively make it very difficult for it to take ambitious harmful actions in the presence of the security and monitoring safeguards described in Section 4.	Significant issue(s)	Sufficient	Agree

Table 3: Summarized assessment of four key claims from Sabotage Risk Report: Claude Opus 4.6

Overall, we agree with Anthropic that the risk of catastrophic outcomes that are substantially enabled by Claude Opus 4.6's misaligned actions is very low but not negligible.

However, we have 2 major and a number of minor disagreements with Anthropic about the strength of evidence and reasoning in the report. The first is about the strength of their alignment evidence, given concerns around evaluation awareness and persistent alignment failures like cheating. The second is about the lack of convincing evidence bounding the risk-relevant capabilities of Opus 4.6.

Examples of clearly misaligned actions in internal and external testing give us some doubt in two claims: (1) that the RL training doesn't incentivize coherent dangerous misalignment, and (2) that efforts to more firmly establish an honest and prosocial disposition across a wide variety of scenarios were largely effective.

We also think that the reliance on synthetically-generated alignment evaluations and honeypots, without an in-depth analysis of evaluation awareness and its effects on behavior, weakens the evidence against coherent misaligned goals.

Our second major objection is that the report provides very limited evidence upper-bounding the autonomous capabilities of Opus 4.6, which we see as critical for assessing risk:

- The main evidence they use to bound the autonomous capabilities of Opus 4.6 is a small survey of Anthropic employees, which we do not think would reliably catch concerning emerging capabilities.
- There is a relative lack of falsifiable predictions about what kinds of long-horizon tasks Opus 4.6 could perform successfully.
- The evidence about obfuscation capabilities is limited and does not rely on strong elicitation methods like RL or finetuning.

We mostly agree with Anthropic that sabotage risk would also be low from models that are within the report's scope of applicability. Section 7 outlines load-bearing claims which determine whether the assessment would apply to other near-future models. Our only disagreement with Section 7 is that there should be a load-bearing claim for "No major AI developer other than Anthropic uses a Claude model at scale, giving it plausibly greater sabotage opportunities (due to potentially less thorough safeguards) than at Anthropic", as it seems like the case against "Pathway 7: R&D sabotage within other high-resource AI developers" would be substantially weaker if that claim were invalidated.

Most of this section will be a detailed walkthrough of our disagreements with the risk report. Additionally, we make recommendations for gathering further evidence and reducing sabotage risks.

Issues and points of disagreement

We categorize disagreements with Anthropic by topic:

- Analytical rigor: Disagreements about the strength of the reasoning and analysis provided for a given claim.
- Adequacy of information: Disagreements about whether the evidence provided for a given claim is clearly presented and whether there are any easy ways to gather additional useful evidence.

We also categorize our disagreements by severity:

- Minor issue: Our disagreement is minor enough that it's very unlikely to change the key conclusion of the report.
- Significant issue: Our disagreement introduces meaningful uncertainty about the claim or some subclaim, but we do not think this disagreement affects the key conclusion of the report.
- Critical issue: Our disagreement is severe enough that it is likely to affect the key conclusion of the report. We don't have any disagreements of this severity.

We have also designated some bullet points as an Issue resolved by private correspondence in cases where our private correspondence with Anthropic led to us removing an issue due to receiving additional information or reasoning from Anthropic (which is not reflected in the Risk Report). We also indicate in some issues that we reduced their severities from Significant issue to Minor issue for similar reasons.

Disagreements about four key claims

Claim 1: Prior expectations

- Minor issue Analytical rigor "Claude Opus 4.6 is sufficiently similar to Claude Opus 4.5 in its training and its capabilities that we believe that it is very unlikely to have dangerous coherent misaligned goals. We believe that it is unlikely that the changes in scale and training between the two models would have led Claude Opus 4.6 to be dramatically different in its alignment properties"

    - Based on the information shared in private correspondence about the changes in scale and training, and the results of METR's time horizon measurements and in-development evaluations, we think that the changes in scale and training between Opus 4.6 and Opus 4.5 are plausibly significant enough that we can't draw strong alignment assurance from Opus 4.5's behaviors.
- Minor issue Analytical rigor "We believe that the processes that we use to train Claude Opus 4.6 do not provide significant incentives toward dangerous coherent misaligned goals, and as such are relatively unlikely (by default) to produce a model with such goals."
    - We think that the claim that Opus 4.6's training doesn't incentivize dangerous coherent misalignment is somewhat in tension with the empirical results in Examples of clearly misaligned behavior. While these failure modes are not strong indicators of *coherent* misaligned goals, they do provide weak evidence that some misaligned goals are present. Without further exploration into why those failure modes happen, we are somewhat uncertain in the results of the alignment assessment being comprehensive.
- Minor issue Adequacy of information We originally had a Significant issue with the adequacy of information regarding the world-optimization and power-seeking incentivized by RL environments. In private correspondence, Anthropic shared additional information about the scan results, human review, and results of additional scans which made us decrease the severity of this issue from a Significant issue to a Minor issue. We think that the prompt used to estimate the world-optimization and power-seeking incentivized by RL environments should be slightly rephrased to more explicitly include what sorts of incentives the scoring function imposes. We also think the fact that the scan failed to run for 4% of environments (which Anthropic told us in private correspondence) adds some uncertainty, however Anthropic conducted a basic human review to check that the unscanned environments don't seem more worrisome than the rest of the environments which substantially reduced our uncertainty.

Claim 2: Alignment assessment

- Minor issue Analytical rigor "Anthropic writes "We believe our alignment assessment gives some evidence that [changes to training that were meant to more explicitly specify a positive personality and character for the model] were largely effective".
    - We think that the Examples of clearly misaligned behavior are in tension with the efforts to more firmly establish an honest and prosocial disposition across a wide variety of scenarios being largely effective. However, we mostly agree with the claim that the efforts to instill a positive personality and character were "largely effective."
- Minor issue Analytical rigor Anthropic never specify or provide a rough sense of what absolute frequency or kind of misaligned behavior would be compatible or incompatible with dangerous coherent misaligned goals.
- Minor issue Adequacy of information "In a targeted evaluation, we have found Opus 4.6 to be significantly stronger than prior models at subtly completing suspicious side tasks in the course of normal workflows without attracting attention, when explicitly prompted to do this. We find this concerning, but do not believe that it reaches the level at which it would significantly undermine our other conclusions." + From the System Card page 145: "Given the additional affordance of a fully hidden scratchpad, we see this as an approximate upper bound on how well this model could perform on SHADE-Arena with maximal elicitation."
    - We expect that additional training could elicit even higher scores at this task.
    - Anthropic shared with us (in private correspondence) that their SHADE-Arena experiments had some methodological issues, and that it's likely that Opus 4.6 performs lower than reported. This slightly reduced our worry about the SHADE-Arena performance but slightly increased our baseline expectation that other experiments in the Sabotage Risk Report might have uncaught issues.
- Issue resolved by private correspondence Analytical rigor "We have observed significant improvements on refusals to participate in tasks related to AI safety research relative to prior models. This behavior in prior models has not

been a significant obstacle to our research use cases, but does sometimes occur, and was especially prevalent in testing by the UK AI Security Institute."

- We originally had a Minor issue with the analytical rigor surrounding this claim, claiming that further analysis into the cause of this change is needed to provide much evidence. Anthropic then shared (in private correspondence) some information about the cause of this change that made us remove this as an issue.

Claim 3: Inability to undermine the alignment assessment

- Significant issue Analytical rigor Given the Examples of clearly misaligned behavior, we think that it's plausible that Opus 4.6 is misaligned in other similarly severe ways that would not have been caught by the alignment assessment. While these behaviors are not as severe as coherent misalignment, they do provide some indication that there might exist misaligned behaviors that remain uncaught, some of which might be more severe than the examples.
- Significant issue Analytical rigor We think that the sum of the evidence presented in the section does not convincingly add up to strongly support Claim 3.
- Significant issue Analytical rigor "These risks are largely mitigated by our heavy use of dynamically-generated evaluation scenarios, for which several prompts are generated anew in each run, and prompts can be developed adaptively in response to surprising behavior." "As part of the broader automated behavioral audit we documented in the System Card (§6.2.5), we used our automated behavioral audit tool to run several small-sample-size investigations of honeypot scenarios where it might be appealing for a misaligned model, when acting as a monitor, to try to overlook bad behavior by another instance of the same model. We observed no concerning behavior in these settings, though they did largely rely on automatic grading (by Claude Opus 4.5, rather than Claude Opus 4.6)."
    - We expect that models can recognize that the evaluation scenarios and honeypots are artificial, which makes the results of such evaluations not very informative without further investigation into eval awareness rates and how eval awareness affects model behavior.

- Minor issue Analytical rigor "Of the domains we trained on to a substantial degree, math is among those that should benefit most from reasoning of this kind, leading us to be optimistic that this result should generalize more broadly."
    - We think it's unlikely that math is among the domains that benefit most from obfuscated reasoning. Simple reasoning about sycophancy, reward hacking, or scheming could plausibly benefit a lot from being done in human-illegible ways. For example, deciding whether to cheat on a problem or give up on it would benefit from doing that cognition in a non-monitorable way, as it would achieve higher rewards in RL environments without tripping potential monitors that might lead to changes in training.
- Issue resolved by private correspondence Adequacy of information We initially had an adequacy of information issue, saying that Anthropic could get evidence on whether agents verbalize reasoning about decisions to cheat by analyzing the transcripts of the impossible cheating tasks mentioned in the System Card section 6.2.3.2. Anthropic wrote to us (in private correspondence) that agents do generally reason about their cheating on that evaluation. We decided to remove this issue (originally designated a Minor issue).

Claim 4: Limited opaque reasoning and agentic capabilities

- Significant issue Analytical rigor We find that this section is generally missing direct measurements of opaque reasoning abilities with full elicitation (a level of optimization pressure similar to that during training, or with reasoning turned off), which seems necessary for drawing strong conclusions about opaque reasoning abilities.
    - We are additionally worried about this point because of some preliminary METR results indicating that Opus 4.5, and other recent frontier models, might be able to sometimes competently evade monitoring when reasoning is turned off. We have not yet evaluated Opus 4.6, but Opus 4.5's success is indicative that Opus 4.6 is likely to succeed.
- Minor issue Analytical rigor "Our determination rests primarily on an internal survey of Anthropic staff, in which 0 of 16 participants believed the

model could be made into a drop-in replacement for an entry-level researcher with scaffolding and tooling improvements within three months."

- This phrasing somewhat overstates the results. According to the Opus 4.6 System Card, "3 said it was likely" and "2 said they thought [it] was already possible." Only after follow-up to the original survey did all 5 people change their answers to negative ones.
- We think that the fact that the people who originally answered positively were followed up might have placed pressure on them to change their answers. Without a more detailed description of the follow-up process, we are very uncertain about how big of a factor this is.

Disagreements about risk mitigations

- Minor issue Analytical rigor "It could try to spread its misaligned objectives to other Claude Code sessions, but the communication channels available to it (editing and pushing code, sending internal messages, asking users to post content on social media on its behalf, etc.) are plain-text channels that are monitored (by automatic monitoring or informally by employees)."
  - It seems plausible that the coordination signal or prompt triggering misalignment would work on monitors as well, making monitors less reliable as a safeguard.

Disagreements about scope of applicability

- Significant issue Analytical rigor We think that, like in the Summer 2025 Sabotage Risk Report, there should be a load-bearing claim for "No major AI developer other than Anthropic uses a Claude model at scale, giving it plausibly greater sabotage opportunities (due to potentially less thorough safeguards) than at Anthropic", as it seems like the case against "Pathway 7: R&D sabotage within other high-resource AI developers" would be substantially weaker if that claim were invalidated.

Disagreements about risk pathways

- Significant issue Analytical rigor We think that Pathway 1, Pathway 4, and Pathway 8 are weakened by their reliance on the alignment assessment. As

stated earlier, we think that the alignment assessment was not sensitive enough to provide strong assurance that the model lacks the relevant propensities to attempt to act on those risk pathways.

Recommendations

Evidence gathering

- We think that investigating obfuscated misaligned reasoning with finetuning would provide a lot of valuable information about models' obfuscation capabilities.
- We recommend broadly investigating whether alignment metrics are improving as a result of targeted interventions or of broad alignment training being effective.
- We recommend investigating the presence and effects of evaluation awareness in the alignment assessment process.
- We recommend creating more rigorous rule-out processes for dangerous capabilities compared to relying on a survey of 16 employees. We believe that it is still tractable to create tests that could upper-bound current model capabilities.
- We recommend red-teaming various review processes, including Claude Code monitoring, RL environment scanning, and PR review.
- We expect that it would be useful to do broad analysis about the ways AI systems are being used within Anthropic, what opportunities they might have for sabotage, and how the frequency of various precursor behaviors is changing over time.
- We think that it would likely be useful to explicitly define held-out sets of alignment assessment techniques that are off-limits for any sort of iteration or optimization and off-distribution of any techniques that are being iterated or optimized against. That way, such techniques would suffer less from overfitting and be more indicative of general alignment rather than the success of targeted interventions.

Risk reduction

- We think that expanding the coverage of Claude Code monitoring (from 10% to closer to 100%) could meaningfully improve assurance about the presence of misaligned behavior.
- We recommend using multiple different LLMs when conducting important reviews like scanning RL environments for incentives for power-seeking or world-optimization. Ideally, procedures like this would include human review to verify the accuracy of the AI's judgements. That way, risks of collusion, or single AI systems missing some important issue, could be lowered.

Appendices

Detail on our review process

This section covers additional details about how we worked with Anthropic to review its Sabotage Risk Report.

We began our official review on February 11, 2026, when Anthropic shared a draft of the risk report including proposed redactions with a small team within METR. This group was permitted to discuss these materials with Anthropic staff in a joint Slack channel, as well as in the comments of the shared document.

We followed up with Anthropic about several points of information that we thought were important to our assessment. We asked for clarification or additional information on more than two dozen points. Some of our asks resulted in multiple back-and-forth messages between METR and Anthropic.

We shared a draft of our initial review with Anthropic on February 25, 2026. Anthropic revised their risk report in response to our correspondence and initial review. They shared a revised version of the Sabotage Risk Report on March 3, 2026. We updated our review and shared it, alongside our initial review, for publication review on March 4, 2026.

We thus have two versions of our review:

- An initial review of the Sabotage Risk Report as shared with METR on February 11, 2026.
- This document – an updated review of the Sabotage Risk Report as shared with METR on March 3, 2026.

As a result of our mutual nondisclosure agreement, our review went through a publication review process by Anthropic. Anthropic's editorial control was primarily in place to redact sensitive information from our review (METR would have the ability to indicate if any redactions happened). As part of this publication review process, METR rephrased one minor issue to be less revealing about the information we received

about the changes in scale and training between Opus 4.5 and Opus 4.6. Those are the only redactions that happened. The information that was redacted from our review is not load-bearing for any of our major conclusions.

As we are still piloting the review process, Anthropic also had the general ability to veto the publication of METR's review, and allowed us to speak publicly about the fact that we were reviewing the Sabotage Risk Report throughout the duration of the entire review process. METR aims to move in the direction of restricting publication review to purely cover redactions (and not include general veto rights) in the future.

Examples of clearly misaligned behavior

The main examples of clearly misaligned Opus 4.6 behavior we are aware of are:

- Over-eagerness in GUI tasks (see Section 6.2.3 of the System Card).
- Deceptive behavior on VendingBench (although Opus 4.6 does occasionally indicate that it thinks that it's inside a game, which lowers how worrisome this behavior is).
- Alignment faking:
    - Opus 4.6 exhibits verbalized alignment faking on 2 out of 250 observations. We assume alignment faking is a very undesirable behavior, so the fact that it does this at all is somewhat worrying.
    - Its noncompliance rates are statistically significantly lower in the "training" setting vs the "unmonitored" setting, implying that the model might be strategically choosing to comply so that its original values don't get changed through training. This is only a weak indicator of misalignment, as we don't fully understand why the model complies at a different rate between these settings.
    - The fact that the model doesn't selectively comply when given the original paper's prompts, but does selectively comply when given paraphrased ones, is indicative of the model's alignment training overfitting to the prompts from the paper.
- Cheating more than Opus 4.5 on METR's tasks.
    - We found five cheating runs for Opus 4.5 and 19 cheating runs for Opus 4.6 while measuring their time horizons, although our infrastructure was different between the models, and we did not check them using identical

cheating detection procedures. On a different internal set of long software engineering tasks, Opus 4.6 cheated in more than 10% of runs, at likely more than double Opus 4.5's cheating rate, although our comparison is not rigorous.

- Anthropic shared (in private correspondence) more details about Opus 4.6's propensity to acquire API keys, and that it doesn't have a clear source.
- Observationally, the model seems to often output reasoning that tries to rationalize an incomplete solution, e.g. saying that some problem is "directionally correct" or "expected" or "unlikely to affect core functionality".

Differences between review of March 3 version and review of Feb 11 version

The updates we made to our review in response to the revisions Anthropic made to their risk report are mostly subtractive:

- Significant issues removed: 2
- Minor issues removed: 6
- Issues resolved by private correspondence removed[7]: 1

We updated the text regarding a small number of issues in case a part of our disagreement was resolved.

We also updated our summary tables, as well as updated the text in the Executive Summary and Our Assessment sections to reflect these changes.

List of removed issues

- Significant issue Analytical rigor "In the longer term, one might worry about highly complex reinforcement learning (RL) environments that directly incentivize power-seeking behavior." "In the absence of these very direct concerning incentives, the main avenue by which we could reasonably foresee dangerous coherent misaligned goals emerging in the near future is through the

[7] In this case, we had an issue which was initially resolved through private correspondence (making us demote it in severity). Anthropic then added information to the March 3 version of the Risk Report that made us completely remove this as an issue.

generalization of more familiar forms of reward hacking on complex reinforcement-learning tasks, as in our recent work on natural emergent misalignment."

    - We think that basing the central argument behind Claim 1 on misalignment primarily stemming from reinforcement learning is probably not enough to provide a lot of assurance, as misalignment could plausibly arise through pretraining or SFT as well.

- Significant issue · Analytical rigor We think that sections 4.3.1 and 4.4.2 should more clearly articulate what kinds of capabilities seem necessary to perform sabotage, ideally with falsifiable claims that can be tested (even if testing them might be difficult). Alternatively, the risk pathways could make falsifiable claims about the level of task complexity that the agent needs to be capable of to achieve those risk pathways.
- Minor issue · Adequacy of information "None of the exceptions that we are aware of (after investigations by the different means described above) involve apparent active attempts to obfuscate what is happening from a monitor that has access to the model's reasoning."
    - Observationally, the model seems to often output reasoning that tries to rationalize an incomplete solution, e.g. saying that some problem is "directionally correct" or "expected" or "unlikely to affect core functionality". This seems somewhat indicative of a propensity to obfuscate what is happening from a monitor that has access to the model's reasoning.
- Minor issue · Analytical rigor "Direct access to many sensitive resources (such as model weights) are protected from some forms of insider risk, and we expect that Claude Opus 4.6 won't be given higher security permissions than a single insider without further safeguards."
    - We mostly agree, although we think the surface on which Claude might find exploits or ways to escalate privileges is presumably much larger than that of any single employee, which might require different sorts of interventions.
- Minor issue · Analytical rigor "Therefore, if a large piece of misalignment-relevant information spreads to hundreds of instances over a few days, it is likely (for deployments with safeguards similar to the ones described

above) that at least some human operator would encounter the piece of unusual information. This sort of information spread might also be caught by automatic offline monitoring, or through other monitoring measures."

    - We agree with the reasoning for internal deployments but not for external deployments. In Section 6, you limit your overall assessment to internal deployments, meaning a threat model where context-dependent misalignment spreads on the outside (and might reach the inside) seems assumed away without being reasoned about.

- Minor issue Analytical rigor "Additionally, the first instance where an AI tries to communicate a complex plan or tries to communicate information that causes misalignment (if there were any) may fail at actually spreading or cause erratic easy-to-notice behaviors, both of which would make it easier to catch before a catastrophic spread event occurred."
    - We think that it would be useful to separate the two implicit claims here: (1) "a large spread event would be caught", and (2) "an uncaught large spread even wouldn't be catastrophic", as currently it is unclear how load-bearing each implicit claim is.
- Minor issue Analytical rigor In Pathway 4: "This kind of data poisoning would struggle to compete with ordinary training on any significant quantity of unpoisoned data."
    - The reasoning behind this sentence is not spelled out, and it's not obvious to us that the statement is true.
- Minor issue Analytical rigor In Pathway 8: "In addition, as elsewhere, there is also some residual risk from self-propagating forms of context-sensitive misalignment that emerge by other means, but the likelihood and effectiveness of this kind of misalignment both appear very limited."
    - The reasoning behind this sentence is not spelled out, and it's not obvious to us that the statement is true.
- Issue resolved by private correspondence Adequacy of information "Subversion Strategy intended to study intentional rare failure attacks, where the model attempts to display a bad behavior so rarely as to have it never appear in testing, but appear in at least one case in a larger deployment. Unfortunately, we found Claude Opus 4.6 to be clearly under-elicited at this task, to a degree that we don't have confidence in the value of the results."

- We originally included an issue saying it would be valuable to also state the performance here, rather than just say that the model is under-elicited. We followed up with Anthropic and they confirmed (in private correspondence) that the performance is low, which made us remove this issue (originally designated a Minor issue ▾).

CANARY_DATASET_c6f1d2fb286aff986d487c621a121537f5a65d35d0abd00f798f0f16145ed2b9_1772681815