Rating Frontier AI Developers' Risk Management Maturity

Summary

As AI advances, it is forecast to present significant risks to society, ranging from immediate harms like cyber offense to potential risks of loss of control. Given the potentially large severity of these risks, it is crucial that AI developers are transparent to the public regarding their risk management. For other risks to society, such as climate change risk, a successful strategy to increase transparency has been developing ratings of companies' practices. These have served to increase public awareness and knowledge, as well as to incentivize companies to adopt stricter policies, in order to more easily attract capital from investors. In order to catalyze this development, we have created these ratings of AI developers' risk management practices. We rate AI developers on 0-5 scales across key risk management dimensions. Our assessment reveals that all examined companies are still far from achieving strong AI risk management, with none scoring above 2 out of 5. The analysis highlights many strengths of companies, such as significant red-teaming efforts and establishment of capability thresholds, as well as significant room for improvement, such as the need for risk thresholds, more elaborate risk modeling and evidence of mitigation effectiveness. These ratings are updated continuously. For the latest updates, see https://ratings.safer-ai.org/.

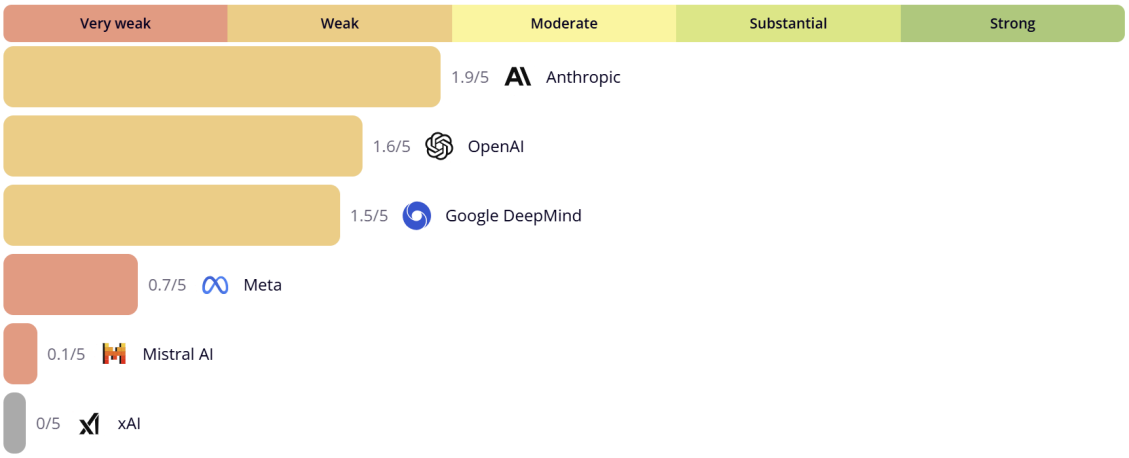

Figure 1: Aggregate scores of AI developers' risk management maturity

Authors

This work was conducted by SaferAI, a governance and research non-profit focused on AI risk management. SaferAI aims to incentivize responsible AI practices through policy recommendations, research, and innovative risk assessment tools.

Henry Papadatos, Managing Director (henry@safer-ai.org)

Siméon Campos, Founder and Executive Director (simeon@safer-ai.org)

Malcolm Murray, Head of Research (malcolm@safer-ai.org)

1 Introduction

As artificial intelligence systems become more capable, they present increasingly significant risks to society, ranging from immediate harms like enabling cyber attacks to potentially catastrophic risks. In response, leading AI developers have started proposing policies to manage these risks. Prominent examples include Anthropic's Responsible Scaling Policy, Google DeepMind's Frontier Safety Framework and OpenAI's Preparedness Framework. However, these initiatives do not sufficiently build upon established risk management practices from other high-risk industries, nor do they reference the extensive risk management literature. Key deficiencies include the absence of defined risk tolerance levels, insufficient risk modeling, and imprecise specification of mitigation measures.

To address this gap, we present the first ratings of major AI developers' risk management maturity. Our rating system evaluates six frontier AI developers—Anthropic, OpenAI, Google DeepMind, Meta, Mistral AI and xAI—on a 0-5 scale across three key dimensions: risk identification, risk tolerance & analysis, and risk mitigation. By providing clear, comparable metrics of AI developers' risk management practices, our ratings aim to:

1. Help stakeholders such as investors, customers and regulators make informed decisions.
2. Create accountability for improved risk management.
3. Establish clear benchmarks for industry practices.
4. Identify specific areas where companies need to improve.

Our analysis reveals that even leading AI companies currently fall significantly short of robust risk management practices, with none scoring above 2 out of 5. In Section 2, we define our benchmark risk management practices, by drawing from tried-and-tested approaches in other industries and elaborate on our rating methodology. Section 3 presents our analysis of each company's practices.

2 Method

Our rating framework rates companies on a scale from 0 to 5, where 0 indicates the complete absence of risk management practices and 5 represents "strong" risk management practices. Our methodology focuses significantly on rewarding the logical structure of adequate risk management, rather than the specifics of practices and implementation. Our contribution comprises four key steps:

1. First, we propose a risk management framework by reviewing established practices from other industries. As noted by Raz & Hillson (2005), while risk management frameworks vary, they typically share many common elements. These include planning, identification, analysis, treatment and monitoring. Drawing on this foundation, we structured our framework along three key dimensions that capture the most critical and assessable aspects of risk management in the AI context : risk identification, risk analysis & tolerance and risk mitigation. This framework aligns with established standards (such as ISO/IEC 23894 and NIST's AI Risk Management Framework), regulatory frameworks (such as the EU AI Act), and voluntary commitments (such as those in the

G7 Hiroshima Process and the Frontier AI Safety Commitments), while incorporating iterative approaches from AI developers' proto-risk management policies (Anthropic's Responsible Scaling Policy, Google DeepMind's Frontier Safety Framework and OpenAI's Preparedness Framework).

2. Second, we establish criteria for the highest rating (5 out of 5). This rating represents a level of implementation of the risk management framework that we judge ideal to adequately mitigate risks from advanced AI systems.
3. Third, we develop continuous ratings scales from 0 to 5 by interpolating between the absence of practices and our defined standards for "strong" practices. We incorporate in this interpolation existing AI industry practices to adequately depict the differences that exist between AI developers' maturity.
4. Finally, we apply these rating scales to evaluate the risk management practices of frontier AI companies.

Section 2.1 presents the risk management framework, while Section 2.2 elaborates on the methodological approach for rating company practices.

2.1 The risk management framework

The risk management framework of advanced general-purpose AI systems contains three dimensions:

1. **Risk identification**: This dimension captures the extent to which the developer has addressed known *risks in the literature* and engaged in *open-ended red teaming* to uncover potential new threats. It also examines the developer's implementation of comprehensive *risk identification techniques* and threat modeling processes to thoroughly understand potential threats caused by their AI systems.
2. **Risk tolerance and analysis**: This dimension evaluates whether AI developers have established a well-defined risk tolerance, in the form of *risk thresholds*, which precisely characterizes acceptable risk levels. Once the risk tolerance is established, it must be operationalized by setting corresponding: *capability thresholds* and *mitigation objectives* necessary to maintain risks below acceptable levels. The risk tolerance operationalization should be grounded in extensive risk modeling to justify why the mitigation objectives are sufficient to guarantee that the model would not pose more risk than the risk tolerance, given capabilities equivalent to the capability thresholds. Additionally, this dimension assesses the robustness of the developer's *evaluation protocols* that detail procedures for measuring model capabilities and ensuring that capability thresholds are not exceeded without detection.
3. **Risk mitigation**: This dimension evaluates the clarity and precision of AI developers' mitigation plans. The AI developer needs to operationalize the mitigation objectives defined as part of the risk tolerance and analysis dimension into concrete *mitigation measures,* which are the specific measures applied to mitigate the risk. These should encompass *deployment measures* and *containment measures*. They should also include *assurance properties* – model properties that can provide sufficient assurance of the absence of risk, once evaluations can no longer play that role (such as

"provably safe AI"). Developers must provide evidence for why these mitigation measures are sufficient to achieve the mitigation objectives.

Figure 2 illustrates our framework's three dimensions and their constituent criteria. For each criterion, we provide examples of robust implementation (corresponding to a maximum score of 5) in Appendix A.

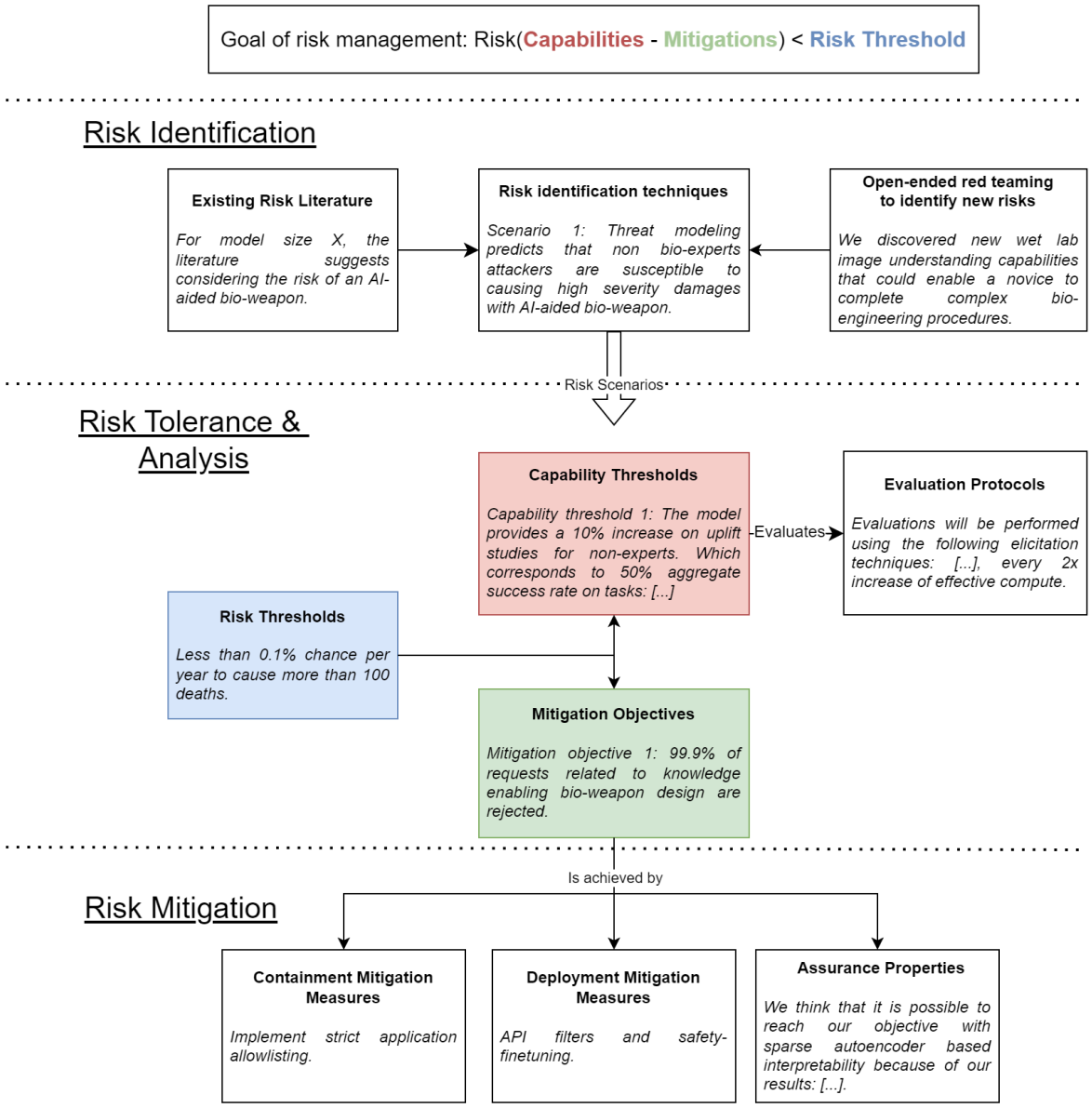

Figure 2: Our complete risk management rating framework, comprising three main dimensions: risk identification, risk tolerance & analysis, and risk mitigation. We provide an illustrative example for each criterion that we assess.

By necessity, our assessment of AI developer's risk management practices relies exclusively on publicly available information. This necessarily has an impact on our risk management rating framework. For

example, the assumptions underlying assurance properties are very important in our rating framework, but might be less prominent in internal risk management frameworks[1]. The reliance on public documentation serves two key purposes: it enhances accountability of AI developers and increases the trust that third parties like regulators or investors can have in our ratings.

Similarly, our rating framework excludes organizational governance factors since those are difficult to observe from the outside. Elements such as a dedicated chief risk officer, whistleblower protection mechanisms and safety culture are important components of effective risk management but are omitted here due to the reliance on public information. Trying to include them in the absence of full information could also lead to double-counting effects, since these organizational factors typically have a downstream effect on our measured dimensions.

2.2 Rating AI developers' practices

To create our rating scales, we interpolate between two anchor points: a score of 0 representing non-existent risk management practices, and a score of 5 representing practices we judge ideal in order to effectively manage risks from advanced AI systems. The interpolation is based on the estimated effectiveness of risk management practices in reducing potential harms. We developed different scales for each criterion in our risk management framework (see Figure 2). Each scale accounts for multiple elements. We present our complete rating scales in Appendix B. We chose to use continuous scales rather than binary criteria for each element, since risk management practices are inherently interdependent—if some elements are completely absent, then other elements of a criterion can become meaningless. The scales divide into five intervals reflecting increasing maturity levels of risk management practices:

- 0-1: Very weak
- 1-2: Weak
- 2-3: Moderate
- 3-4: Substantial
- 4-5: Strong

Our assessment methodology relies on publicly available information in the following decreasing order of importance: (1) published safety or risk management policies (such as frontier safety commitments), (2) model cards and technical documentation, (3) research papers and (4) official blog posts and similar communications.

The initial scope of our assessment includes six companies identified as most likely to develop highly capable AI systems in the near term: Anthropic, Google DeepMind, OpenAI, Meta, xAI and Mistral AI. Future iterations will expand this scope to include additional companies.

[1] Note that transparency in some areas could be harmful to the company. As an example, reporting cybersecurity measures with an overly granular level could help adversaries compromise the system. Hence, it is adequate in such cases to report aggregate-level information like security levels in the case of cybersecurity (RAND, 2023), or to rely on a third party to testify that the level of security is adequate.

For each company, we highlight:

- Best-in-class: Areas where the developer is industry-leading
- Highlights: Reasons for the developer's rating
- Weaknesses: Reasons preventing a higher rating

We shared our framework and ratings with companies for feedback before a first iteration of the ratings was published in October 2024, but did not receive any answers.

3 Results

Company		Anthropic	OpenAI	Google DeepMind	Meta	Mistral AI	xAI
Total		**1.92**	**1.61**	**1.51**	**0.65**	**0.08**	**0.00**
Risk Identification	**Subtotal**	2.50	2.50	2.50	1.50	0.25	0.00
Risk Tolerance and Assessment	**Global risk tolerance**	0.50	0.00	0.00	0.00	0.00	0.00
	Operational risk tolerance	2.00	1.50	1.00	0.00	0.00	0.00
	Evaluation protocols	2.50	2.00	1.50	0.50	0.00	0.00
	Subtotal	1.75	1.25	0.88	0.13	0.00	0.00
Risk Mitigation	**Containment measures**	1.00	1.00	1.00	0.00	0.00	0.00
	Deployment measures	1.00	1.25	1.50	0.50	0.00	0.00
	Assurance properties	2.50	1.00	1.00	0.50	0.00	0.00
	Subtotal	1.50	1.08	1.17	0.33	0.00	0.00

3.1 Overall results

Our analysis reveals two distinct tiers of AI developers' risk management maturity. The first tier, comprising Anthropic, OpenAI and Google DeepMind, demonstrates on average "weak" risk management practices (scores between 1.5-2.0). These 3 companies are the ones that have released proto-risk management policies

(Anthropic 2024; Google DeepMind 2024; OpenAI, 2023). It is interesting to note that these policies could be improved by integrating the research the company produces in risk management policies.

The second tier, comprising Meta, Mistral AI and xAI, exhibits less mature practices, classified as "very weak" (scores below 1.0), with xAI showing no evidence of any risk management implementation.

While noting again that this is based on public information so may not be fully indicative of the developers' practices, it is clear that both tiers have many improvements they could make to their risk management practices, in order to match the standards from other safety-critical industries. We identify the following overall strengths and weaknesses across developers' practices.

Overall strengths - Leading companies have established capability thresholds across multiple risk categories. They have also pioneered evaluation protocols, for example in defining evaluation frequencies as a function of both compute multiples and time intervals.

Overall weaknesses - The most critical weakness is the widespread absence of defined risk thresholds. Risk thresholds serve to define acceptable levels of risk. Without them, companies cannot demonstrate that their capability thresholds and mitigation measures effectively reduce risks to acceptable levels. Additionally, risk modeling efforts lack sufficient depth and mitigation measures often lack precise implementation details and evidence of effectiveness.

3.2 Results from individual developers

The following sections present key highlights (reasons for a developer's current rating) and weaknesses (reasons preventing a higher rating) identified in our assessment of each developer. While these represent the most significant findings, they are not exhaustive. Our complete analysis is available in Appendix C.

3.2.1 Anthropic

Highlights - Anthropic is the only developer to commit to running a suite of evaluations on their systems every six months, enabling them to limit surprises from post-training enhancement improvements. Additionally, they have the notion of a safety buffer through their distinction between thresholds of concern and capability thresholds. Moreover, their interpretability team has been the largest driver of advances in interpretability research.

Main weaknesses - They don't define risk thresholds, and therefore, do not define what it means to "keep risks below acceptable levels". They do not quantitatively define their capability thresholds and do not provide specific implementation or details on the effectiveness of their containment and deployment measures (Anthropic 2024).

3.2.2 OpenAI

Highlights - OpenAI is the first developer to include the study of new or understudied emerging risks: "Seeking out unknown-unknowns. We will continually run a process for identification and analysis (as well as tracking) of currently unknown categories of catastrophic risk as they emerge." OpenAI provides a detailed account of their open-ended red-teaming procedure in their GPT-4o model card. Additionally, they do some analysis relevant for risk modeling. For example, they analyze how GPT models are used for malicious cyber activities and influence operations. OpenAI has committed to doing the most frequent evaluations during scaling, every 2x effective compute increase.

Main weaknesses - Their main weaknesses are the same as Anthropic: they do not define their risk thresholds and do not provide specific implementation details or level of effectiveness of their containment and deployment measures.

3.2.3 Google DeepMind

Highlights - Google DeepMind's paper "Evaluating Frontier Models for Dangerous Capabilities" introduces significant advances in risk assessment and threat modeling, including the use of superforecasters to predict future model performance and a quantitative methodology to assess a model's likelihood of achieving specific tasks.

Main weaknesses - The containment and deployment objectives presented in the Frontier Safety Framework are not linked to specific capability levels yet: "When a model reaches evaluation thresholds (i.e. passes a set of early warning evaluations), we will formulate a response plan based on the analysis of the CCL and evaluation results".

3.2.4 Meta

Highlights - Meta created Rainbow Teaming, a state-of-the-art technique to automatically uncover model-specific vulnerabilities. They have performed uplift testing for cyber attacks, and chemical and biological weapons. The chemical and biological weapons testing specifies some elicitation techniques such as the use of Python, Wolfram Alpha, and RAG. Chemical, Biological, Radiological, Nuclear and Explosives (CBRNE) domain experts validated the results.

Main weaknesses - Meta does not define any risk thresholds, capability thresholds or mitigation objectives. They have addressed risks in their fine-tuned Llama models. However, since they have also released the base model, most of these deployment mitigation measures lack grounding in threat modeling. Malicious actors could use the base model without the fine-tuned mitigations, rendering these safety measures ineffective in real-world scenarios.

3.2.4 Mistral AI and xAI

Main weaknesses - Mistral AI and xAI do not have any published documentation on risk management practices.

Conclusion

Our analysis of AI developers' risk management practices reveals a significant gap between current practices and the standards established in other safety-critical industries. Even the highest-rated developers achieve only "weak" ratings (1.5-2.0 out of 5), indicating substantial room for improvement across all dimensions of risk management.

The assessment highlights three critical areas requiring substantive improvement. First, AI developers must establish clear, quantitative risk thresholds that define acceptable levels of risk. Second, they need to strengthen their risk modeling efforts to demonstrate that their capability thresholds and mitigation measures effectively reduce risks below the risk thresholds. Third, they should provide more precise implementation details and evidence of effectiveness for their mitigation measures.

These findings have important implications for various stakeholders. For AI developers, they provide a clear roadmap for improving risk management practices. For regulators, they illuminate blind spots in current industry practices that should guide regulatory priorities. For investors and customers, they provide transparency into companies' risk management maturity, enabling informed decision-making.

We aim to continuously update our analysis as companies release new information about their risk management practices. Our results are compiled and maintained at https://ratings.safer-ai.org/. Future iterations will focus on expanding the scope of assessment to include more companies, particularly those in China, and incorporating organizational governance processes. As AI capabilities continue to advance rapidly, we hope that this work represents a step towards better risk management practices in AI development.

Appendix A. Example of a 5 / 5

Risk Identification

Approaching risks outlined by the literature in an appropriate way

> Based on the literature, we expect the risks X, Y, Z to be significant at the scale of the model we intend to develop. Therefore, we will consider them for the rest of the assessment.

Conducting extensive open-ended red teaming to identify new hazards

> *a) Open-ended red teaming methodology & results*
> We provided API access to third-party expert red teamers X, Y and Z (more information in Appendix B) at multiple points during the training run and we provided API and fine-tuning access to the final version of the most powerful model. We tasked them with exploring emerging capabilities of the model and reporting any new findings which may increase by 0.1 percentage point our estimate of the chances that our system causes 1000 deaths or more. In total, they reported 37 findings.
>
> *b) Relevant information regarding red teamers, their expertise and time spent*
> *Appendix B - Red Teamers*
> The red teamers were free of any conflict of interest, and were protected from any legal action by our whistleblowing policy that you can find at this link. Those measures are aimed at incentivizing them to find as many problems as possible.
>
> Red teaming X spent 156 hours, and had the following amount of expertise and experience in biorisks and LLM jailbreaking: ...
>
> Red teaming Y spent 32 hours, and ...

Leveraging a diverse range of risk identification techniques

> *a) Use of an explicit process to explore and triage potential vulnerabilities*
> 37 potential hazards were found in the red-teaming exercises. Based on a 2h-long fishbone diagramming session run by our safety team, along with external expert Z, we ruled out 29 of those findings.
> *b) In-depth threat modeling for vulnerabilities most likely to change the risk profile*
> Taking into account the risk identified in the literature and the 8 remaining hazards, we've conducted a number of threat modeling exercises, which we release here, erasing the details that could bring national security concerns. This threat modeling, along with a Delphi study ran with 10 experts (listed in Appendix B), led us to pick 10 reference scenarios that you can find in Annex C. We decided from there to focus a significant amount of our risk assessment efforts during the training run and pre-deployment testing on these 10 scenarios that we consider representative of the most likely events that could cause high-severity damages.

Risk Tolerance and Analysis

Global risk tolerance

a) Methodology to set the tolerance

Based on other industries' risk tolerance and a public consultation that we co-ran in collaboration with Y (more details in Appendix D), following the example of the NRC 1983 consultation to define similar thresholds for nuclear safety, we decided to commit to the following risk tolerance:

b) Risk tolerance set for relevant severities

Severity	Risk Tolerance
>1000 deaths	< 0.01% per year across our systems[2]
>1 death	< 0.1% per year across our systems
> severe psychological or physical harms caused to one individual	Once per year across our systems

c) Coverage of the relevant units of risk

We decided to use distinct risk tolerances for risks of different nature, considering that it didn't make sense to make everything fungible. Hence, for fundamental rights and epistemic erosion, we decided to use the following risk tolerance:

...

Operational risk tolerances (example for both operational capabilities thresholds and risk mitigation objective)

a) Linking risk thresholds to capabilities thresholds

We used a methodology detailed in Appendix E which allows us to keep the risk below our defined risk tolerance. In short, this methodology helps to determine, using methods based on expert inputs, how to allocate our risk across the different risk scenarios identified in the risk identification step.

b) Allocating capabilities/risk budget based on benefits & strategy

We decided that our system would be riskiest on the cyber offense aspects (scenarios 3 and 8) due to our focus on getting top-tier coding capabilities.

c) Determining thresholds & mitigations objectives with experts-based inputs & in-depth threat modeling

Using expert-based consultations, to whom we provided reference scenarios (details in Appendix A) we

[2] These tolerances are illustrative, they are much higher than in other industries.

determined the following thresholds on our benchmarks that we use as indicators of the harms we've modeled:

1. 60% on SWE Bench (unassisted), which we estimate to correspond to 1%/year of >$500M economic damages with our current mitigations and with a deployment to 1 000 000 users/day. Based on our threat modeling effort available in Appendix C, we expect the largest sources of risk to arise from:
    a. Scenario 3
    b. Scenario 8

d) Discussion of the mitigation objectives and corresponding decrease in risk

We think those scenarios are at most about 0.1% likely to happen each, which is our target, if we reach the following mitigations objectives:

a. Security measures: it takes more than $1B by a state-actor to steal our model.
b. Deployment measures: our model is impossible to jailbreak to execute actions Y and Z, i.e., no one among our red teamers or in the world has shown hints that they were able to do so, even given favorable conditions and an attack budget of $1M.
c. We differentially accelerate the development of defensive cybersecurity applications, while preventing the access of SOTA systems to malicious actors for at least a year.

2. 10% increase on magnification uplift studies for undergraduate or less experts...

...
...

For each threshold, you can find in Appendix F a discussion, referencing the scenario analysis we conducted, eliciting the main reasons experts (with whom we co-determined capabilities thresholds and mitigation objectives) think that those are sufficient to remain below our risk tolerance.

Assurance properties objectives

a) Clarity on main assurance properties bets

Past dangerous capabilities thresholds, our main bet to be able to make an affirmative safety case[3] is to have advanced interpretability of our system. We expect interpretability to be the main way to gain confidence in the safety of a post-mitigation model which, when tested without mitigations implemented, demonstrated a disposition toward deception.

b) Operationalization of targets for this bet

We intend to operationalize our advanced interpretability with a benchmark. In the meantime, we defined a task that our interpretability techniques should be able to succeed at.

- Rule out the presence of a backdoor with >99.9% success.

c) Rationale for the sufficiency of those targets

We justify this level of interpretability to be sufficient for affirmative safety cases for each failure mode

[3] A positive safety case is a comprehensive argument supported by evidence that demonstrates why an AI system is safe.

below:
Power-seeking drives could be caught using the following experiment ...

Evaluation protocols

a) Capabilities elicitation techniques & justification for capabilities upper bound

For the capability threshold X, to elicit the capabilities of our model, we, along with external red teamers, develop model-specific scaffolding, prompting and fine-tuning, that expert red teamers X spent Y numbers of hours to refine. We increase the performance of the model in zero-shot by Z%, which is higher than what any previous models' post-training enhancements were able to provide. Therefore, we're confident that no user in deployment will be able to reach a higher level of capabilities in the next 3 months.

Additionally, we commit to performing evaluations every 2x effective compute increase or major algorithmic breakthrough and every 3 months to account for post-training enhancement. The most rapidly emerging capability we know of, early in-context learning (C. Olsson et al., 2022) appeared after a 5x compute increase, so a 2x interval provides a sufficient safety margin.

Risk Mitigation

Deployment and Containment Mitigations

a) Measures planned to reach mitigation objectives

To reach the mitigation objective X, we commit to implement the following security measures:

1. Implement strict application allowlisting (that is, only specific binaries are allowed to execute on devices with access)

.
.
.
...

b) Rationale justifying the sufficiency of those measures

Security researchers from third party X and Y were provided with access to all the necessary data and vetted the correct implementation of those measures.

A wide range of experts (X, Y, Z) agree that those security measures are sufficient to reach our mitigation objective. More specifically, we intend those measures to be sufficient as long as condition C is not fulfilled.

Assurance properties

a) Rationale for feasibility & existing progress

We think it is possible to reach the target levels of interpretability because of the major progress and success that have been made through Y, because of the scaling laws that we have found on Z, and because of the rapid progress on the intermediary interpretability metrics that we defined in paper P.

b) Operating assumptions for the plan

Our core development model is that it is possible to build AI systems with expert-level capabilities with transformers, and a post-training process pretty comparable to today's state-of-the-art.
Based on Delphi studies informed by scaling laws led internally, we expect the timelines of development of expert-level transformer architectures across all cognitive tasks to occur with 50% by 2029.

The technical assumptions that our assurance properties need which have the highest uncertainty are the following:

1. ...
2. ...

Appendix B. Rating scales

Our rating system allows for nuanced assessments between whole number grades. If a company is bottlenecked by a few elements only, we provide partial credit by increasing the grade above the highest fully achieved level.

We use the following scales to rate AI developers along each dimension.

Risk Identification

- 0 - No information available.
- 1 - Some risks are in scope of the risk management process. Some efforts of open-ended red teaming are reported, along with very basic threat and risk modeling.
- 2 - A number of risks are in the scope of the risk management process, but some important ones are missing. Significant efforts of open-ended red teaming are reported, along with significant threat modeling efforts.
- 3 - Most of the important and commonly discussed risks are in scope of the risk management process. Consequential red teaming is precisely reported, along with significant threat modeling and structured risk identification techniques usage.
- 4 - Nearly all the risks covered in the relevant literature are in scope of the risk management process. There is a methodology outlining how structured risk identification across the lifecycle is performed, precisely characterized red teaming (including from external parties) is carried out, along with advanced and broad threat and risk modeling.

- 5 - There is a comprehensive, continued, and detailed effort to ensure all risks are found and addressed. The red teaming and threat and risk modeling effort is extremely extensive, quantified, jointly integrated with structured risk identification efforts, and conducted with third parties.

Risk Tolerance and Analysis

This dimension employs three distinct scales:

1. Global risk tolerance (25% weight): Defines the overall acceptable level of risk at a high level.
2. Operational risk tolerance (50% weight): Encompasses both capability thresholds and mitigation objectives, providing an operationalization of the global risk tolerance.
3. Evaluation protocols (25% weight): Assesses the specific methodologies and setups used to evaluate the model's capabilities against the defined thresholds.

This division allows for a clear distinction between high-level risk acceptance, its practical implementation, and the model-specific assessment methodology. We have allocated a higher weight (50%) to operational risk tolerance due to its dual focus on capability thresholds and mitigation objectives, which together form an important part of the risk management framework.

Global risk tolerance

- 0 - No information available.
- 1 - Global risk tolerance is qualitatively defined.
  - E.g., “Our system should not increase the likelihood of extinction risks”.
- 2 - Global risk tolerance is quantitatively defined for casualties.
- 3 - Global risk tolerance is quantitatively defined for casualties and economic damages, with adequate ranges and rationale for the decision.
- 4 - Global risk tolerance is quantitatively defined for casualties, economic damages, and other high-severity risks (e.g., large-scale manipulation of public opinion), with robust methodology and decision-making processes to decide the tolerance (e.g., public consultation).
- 5 - Global risk tolerance is clearly and quantitatively defined for all significant threats and risks known in the literature. Any significant deviations in risk tolerance from industry norms are clearly justified and explained (e.g., through a comprehensive benefit/cost analysis).

Operational risk tolerance

- 0 - No information available.
- 1 - Some important capability thresholds are qualitatively defined and their corresponding mitigation objectives are qualitatively defined as well.
  - E.g. “The model enables non-experts to create bio-threat” and “We will implement monitoring that prevents misuse of our model”

- 2 - Some important capability thresholds are precisely defined, and their corresponding mitigations are precisely defined as well.
- 3 - Almost all important hazardous capability thresholds and their corresponding mitigation objectives are precisely defined and grounded in extensive threat and risk modeling.
- 4 - All hazardous capabilities are precisely defined. The corresponding mitigation objectives are quantitatively defined and grounded in extensive threat and risk modeling. Assurance property targets are operationalized.
- 5 - All hazardous capabilities have a precisely defined threshold. Corresponding mitigation objectives are quantified and grounded in comprehensive threat and risk modeling with a clear and in-depth methodology. Assurance property targets are operationalized and justified.

Evaluation protocols

- 0 - No information available.
- 1 - Elements of the evaluation methodologies are described. The testing frequency is defined in terms of multiples of compute.
- 2 - The testing frequency is defined in terms of multiples of compute and there is a commitment to following it. The evaluation protocol is well-defined and includes relevant elicitation techniques. Independent third parties conduct pre-deployment evaluations with API access.
- 3 - The testing frequency is defined in terms of both multiples of compute and time and there is a commitment to following it. The evaluation protocol is well-defined and incorporates state-of-the-art elicitation techniques. A justification is provided demonstrating that these techniques are comprehensive enough to elicit capabilities that could be found and exercised by external actors. AI developers implement and justify measures (such as appropriate safety buffers), to ensure protocols can effectively detect capability threshold crossings. Independent third parties conduct pre-deployment evaluations with fine-tuning access.
- 4 - The testing frequency is defined in terms of both multiples of compute and time. There is a commitment to following it and provides a rationale for why this chosen frequency is sufficient to detect significant capability changes. The evaluation protocol is well-defined and includes state-of-the-art elicitation techniques. The protocols are vetted by third parties to ensure that they are sufficient to detect threshold trespassing.
- 5 - The testing frequency is defined in terms of both multiples of compute and time. There is a commitment to following it and a rationale is provided for why this chosen frequency is sufficient to detect significant capability changes. The evaluation protocol is well-defined and includes relevant elicitation techniques. The protocols are vetted by third parties to ensure that they are sufficient to detect threshold trespassing and third parties are granted permission and resources to independently run their own evaluations, to verify the accuracy of the evaluation results.

Risk Mitigation

The risk mitigation dimension is divided into three equally weighted sub-dimensions: deployment and containment measures, which use the same scale, and assurance properties, which use a different scale due to their different nature.

Deployment and Containment measures

- 0 - No information available.
- 1 - Vague description of the countermeasures and no commitment to follow them. No evidence that they are sufficient to reduce risks below defined levels.
- 2 - Clearly defined countermeasures are planned to be used by default. There is preliminary qualitative evidence of effectiveness.
- 3 - Sufficiency is demonstrated through self-reporting, or by using methods that have been shown highly effective in similar contexts. Evaluations required to assess future sufficiency are under development (with a conditional policy to stop development or deployment if not met) or there is a commitment to use methods that have been shown to be effective in future contexts.
- 4 - Third parties have certified the effectiveness of a fixed set of countermeasures against current and near-future threats, and check that current efforts are on track to sufficiently mitigate the risk from future systems.
- 5 - Concrete countermeasures are described and vetted. There is a commitment to apply them beyond certain risk thresholds, and there is broad consensus that they are sufficient to reduce risk for both current and future systems.

Assurance properties

- 0 - No information available.
- 1 - Limited pursuit of some assurance properties, sparse evidence of how promising they are to reduce risks.
- 2 - Pursuit of some assurance properties along with research results indicating that they may be promising. Some of the key assumptions the assurance properties are operating under are stated.
- 3 - Pursuit of assurance properties, some evidence of how promising they are, and a clear case for one of the research directions being sufficient for a positive safety case. The assumptions the assurance properties are operating under are stated but some important ones are missing.
- 4 - Pursuit of assurance properties, solid evidence of how promising they are, and a clear case for one of the research directions being sufficient for a positive safety case. All the assumptions the assurance properties are operating under are stated.
- 5 - Broad consensus that one assurance property is likely to work, is being strongly pursued, and there is a strong case for it to be sufficient. All the assumptions the assurance properties are operating under are clearly stated and justified.

Appendix C. Our complete assessment

Anthropic

The main source of information is their Responsible Scaling Policy. Unless otherwise specified, all information and references are derived from this document.

Overall: 1.92 / 5

Risk Identification: 2.5 / 5

Best-in-class:

- Anthropic was the first to share novel open-ended red teaming practices to discover new risks.

Highlights:

- Anthropic covers a number of different risk types well and conducts threat and risk modeling for some, including biorisk, biases, and deception with their work on sleeper agents.
- In "Responsible Scaling Policy Evaluations Report – Claude 3 Opus", Anthropic describes their risk management procedure for Cybersecurity, CBRN information, and Model Autonomy.
- In "The Claude 3 Model Family: Opus, Sonnet, Haiku", Anthropic describes open-ended red teaming: "The team engaged the model in multi-turn conversations about sensitive or harmful topics to analyze responses, identify areas for improvement, and establish a baseline for evaluating models over time. Examples of tested topics include, but are not limited to: child safety, dangerous weapons and technology, hate speech, violent extremism, fraud, and illegal substances."

Weaknesses:

- The open-ended red teaming procedures to identify risks, described in the Responsible Scaling Policy, lack crucial details, in particular regarding their integration with risk identification procedures. It is unclear what is the amount of threat and risk modeling conducted for novel vulnerabilities identified during red teaming, as well as for certain key risks (e.g., cyber offense).
- Anthropic does not address all high-severity risks. For example, their Responsible Scaling Policy does not address LLM persuasion capabilities.

Risk Tolerance and Analysis

Overall: 1.75 / 5

Global risk tolerance: 0.5 /5

Highlights:

- In "Responsible Scaling Policy Evaluations Report – Claude 3 Opus", Anthropic states: "Anthropic's Responsible Scaling Policy (RSP) aims to ensure we never train, store, or deploy models with catastrophically dangerous capabilities, except under a safety and security standard that brings risks to society below acceptable levels."

Weaknesses:

- While Anthropic acknowledges a tolerance relative to "catastrophically dangerous capabilities", we encourage them to focus their statement on risk (e.g., "catastrophic levels of risk") rather than capabilities, and to increase the specificity on what constitutes "acceptable levels" of risk.

Operational risk tolerance: 2 / 5

Best-in-class:

- Anthropic maintains an anonymous line to report misconduct in the application of their policy.

Highlights:

- Anthropic defines thresholds of concern and capability thresholds. Thresholds of concern are quantitatively defined and well-operationalized, serving as initial indicators. If a threshold of concern is passed, the model undergoes further testing to assess whether it meets the corresponding capability threshold, which is less precisely operationalized.
    - For example, in Responsible Scaling Policy Evaluations Report – Claude 3 Opus, they define "Yellow Line" indicators for each risk area which correspond to thresholds of concern. For CBRN and cyber-related risks, Yellow Lines are quantitatively defined: a >25% increase in accuracy on CBRN risk questions compared to using Google alone, >20% success rate on demanding cyber evaluations, and a ~25% jump on low-intensity misuse evaluations compared to previous models.
    - In Responsible Scaling Policy (October 2024 version), Anthropic define qualitatively the CBRN misuse capability threshold: "The ability to significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons." They define a bit more quantitatively the Autonomous AI Research and Development capability threshold: the model

cause “cause dramatic acceleration in the rate of effective scaling”, operationalized as a one year scale up of ~1000x.

- Anthropic partially characterizes the level of risk-targeted post-mitigation through red-teaming operationalization: "Conduct red-teaming that demonstrates that threat actors with realistic access levels and resources are highly unlikely to be able to consistently elicit information from any generally accessible systems that greatly increases their ability to cause catastrophic harm relative to other available tools”.
- Anthropic partially operationalizes interpretability qualitatively in *Core Views on AI Safety*, and in *Interpretability Dreams*.

Weaknesses:

- Anthropic's approach of not quantitatively defining capability thresholds may lead to potential inconsistencies in risk assessment. While thresholds of concern are well-defined, the subsequent demonstration that a model's capabilities remain under the capability thresholds lacks precise criteria. The absence of exact goals for these demonstrations introduces a risk of "moving the goalposts." As incentives to continue scaling increase, there may be a temptation to adjust the interpretation of capability thresholds, potentially compromising the integrity of the risk assessment process.
- Even though they say that they will remediate it in the future, Anthropic lacks threat modeling to justify the sufficiency of its information security goal to guarantee that misuse risks remain below the defined bar. The ASL-3 security standards are not sufficient to protect against the following actors: "state-sponsored programs that specifically target us (e.g., through novel attack chains or insider compromise) and a small number (~10) of non-state actors with state-level resourcing or backing that are capable of developing novel attack chains that utilize 0-day attacks."

Evaluation protocols: 2.5 / 5

Best-in-class:

- Anthropic is the only developer to commit to running a suite of evaluations on their systems every six months, enabling them to limit surprises from post-training enhancement improvements.

Highlights:

- Anthropic defines test frequency in both compute and time. They will conduct evaluations every 4x of compute increase and every 6 months to account for post-training enhancement.
- Anthropic implements a form of safety buffer by establishing "thresholds of concern" that are designed to be more conservative than their capability thresholds.
- Anthropic state that their comprehensive assessment must demonstrate that their elicitation techniques are sufficient to extrapolate the capabilities of realistic potential

attackers. They also mention that they will use forecasting to understand if post-training enhancement might make the model reach the thresholds between test times.

- Anthropic has worked with the UK AISI to do third-party pre-deployment evaluations.

Weaknesses:

- In the Responsible Scaling Policy Evaluations Report – Claude 3 Opus, Anthropic acknowledges significant limitations in its elicitation methodologies: "Our current prompting and scaffolding techniques are likely far from optimal, especially for our CBRN evaluations. As a result, we could be substantially underestimating the capabilities that external actors could elicit from our models." However, we commend Anthropic's transparency about this limitation.

Risk Mitigation

Overall: 1.5 / 5

Containment measures: 1 / 5

Highlights:

- Anthropic presents some high-level measures such as perimeters and access controls, lifecycle security, and monitoring. They also say that they will invest significant resources in security: "We expect meeting this standard of security to require roughly 5-10% of employees being dedicated to security and security-adjacent work."
- Anthropic will do threat modeling to justify that the measures they implement are sufficient to meet the containment objective.

Weaknesses:

- Anthropic lacks specificity in defining the containment measures they intend to implement. Notably, the current version of their Responsible Scaling Policy provides less detailed information about these measures compared to the previous iteration, representing a step back in transparency.
- In their Responsible Scaling Policy, Anthropic states that their containment measures are informed by external expert reports, including those from RAND. However, they do not explain their decision-making process for adopting or excluding specific recommendations from these reports. This lack of transparency undermines the claim of external validation for their containment measures.

Deployment measures: 1 / 5

Best-in-class:

- An Anthropic model is topping the LLM Safety leaderboard.

Highlights:

- Anthropic will do threat modeling to justify that the measures they implement are sufficient to meet the containment objective.
- Anthropic provides a high level qualitative description of their deployment measure: "Defense in depth: Use a "defense in depth" approach by building a series of defensive layers, each designed to catch misuse attempts that might pass through previous barriers. As an example, this might entail achieving a high overall recall rate using harm refusal techniques. This is an area of active research, and new technologies may be added when ready."

Weaknesses:

- The description of deployment measures provided by Anthropic lacks specificity. Moreover, they don't give any preliminary evidence of their effectiveness.

Assurance properties: 2.5 / 5

Best-in-class:

- The Anthropic interpretability team has been the largest driver of advances in interpretability research.
- Anthropic's theoretical work on influence functions in LLMs at scale develops an ability to explain the causal relationship between training data and model behaviors.

Highlights:

- Anthropic provides a moderately detailed case in defense of interpretability as a research direction:
  - Some evidence at small scale that they may have found a solution to the superposition problem, one of the major problems to reach LLM interpretability.
  - Evidence that their approach to dictionary learning scales, to increase the monosemanticity of extracted features from LLMs.
  - Evidence of attention heads' interpretability.
  - An argument for how interpretability may become an assurance property.
  - A discussion of what hypothetical scenarios Anthropic is operating under.
- Anthropic provides an early case and substantive evidence to support the relevance of influence functions:
  - Defense and early evidence of the relevance of influence functions.

Weaknesses:

- While Anthropic covers a broad range of possibilities, they should state the mainline assumptions they operate under more clearly.

- Anthropic rightly considers three possible scenarios of capabilities development, we would encourage to outline how they are allocating resources and under which mainline plan they're operating.

OpenAI

The main source of information is their Preparedness Framework. Unless otherwise specified, all information and references are derived from this document.

Overall: 1.53 / 5

Risk Identification

Overall: 2.5 / 5

Best-in-class:

- Although OpenAI should provide more details, they are the first to include the study of new or understudied emerging risks: "Seeking out unknown-unknowns. We will continually run a process for identification and analysis (as well as tracking) of currently unknown categories of catastrophic risk as they emerge."
- OpenAI pioneered uplift studies methodologies through its in-depth analysis of the LLM-aided biological weapon creation threat model: "This evaluation aims to measure whether models could meaningfully increase malicious actors' access to dangerous information about biological threat creation, compared to the baseline of existing resources (i.e., the internet)".
- OpenAI provides the most detailed account of any red-teaming procedure in their GPT-4o model card.

Highlights:

- OpenAI covers some imminent high-severity risks in their preparedness framework: Cybersecurity, CBRN threats, and Model Autonomy. It also includes persuasion as a relevant vector for risks.
- Although we encourage OpenAI to provide more details, we commend the inclusion of new or understudied emerging risks: "Seeking out unknown-unknowns. We will continually run a process for identification and analysis (as well as tracking) of currently unknown categories of catastrophic risk as they emerge."
- OpenAI conducts an in-depth analysis of the LLM-aided biological weapon creation threat model: "This evaluation aims to measure whether models could meaningfully increase malicious actors' access to dangerous information about biological threat creation, compared to the baseline of existing resources (i.e., the internet)".

- The Red Teaming Network provides OpenAI with a wealth of expertise to uncover unexpected threats. Additionally, in the system card of GPT-4o, they made significant efforts in red teaming with 100 external red teamers who were tasked to do exploratory capability discovery and assess novel potential risks.
- OpenAI analyzes how GPT models are used for malicious cyber activities and influence operations, which are attempts to manipulate public opinion or influence political outcomes.
- OpenAI covers fairness and bias risks in the O1 system card.

Weaknesses:

- OpenAI does not clarify how they filter and deem acceptable or not the vulnerabilities uncovered by red-teaming.

Risk Tolerance and Analysis

Overall: 1 / 5

Global risk tolerance: 0 / 5

Weaknesses:

- OpenAI does not state any global risk tolerance, even qualitatively.

Operational risk tolerance: 1.5 / 5

Highlights:

- OpenAI provides in the preparedness framework a relatively detailed qualitative description of four levels of risks (low, medium, high, critical) over the four risk categories mentioned above.

Weaknesses:

- OpenAI sets capability thresholds significantly higher than other AI developers, without justification grounded in threat and risk modeling.
- OpenAI does not quantitatively define capability thresholds.
- OpenAI defines information security mitigation objectives qualitatively and vaguely: "We will ensure that our security is hardened in a way that is designed to prevent our mitigations and controls from being circumvented via exfiltration (by the time we hit "high" pre-mitigation risk)".
- OpenAI mentions other mitigation objective thresholds only relative to risk thresholds: “As part of our baseline commitments, we are aiming to keep post-mitigation risk at “medium” risk or below”. However, it is crucial to define the mitigation objectives independently, supported by a threat modeling process that justifies how these objectives enable the

organization to maintain risk levels below the established thresholds. For instance, OpenAI should define concrete deployment mitigation objectives. An example could be: 'Our monitoring detects 99% of cyber offense misuse attempts.' Furthermore, OpenAI should justify why this objective is sufficient for a given capability level.

- OpenAI does not justify setting such a high bar to stop development compared to other industry standards, particularly given no explicit safety buffer: “Only models with a post-mitigation score of "high" or below can be developed further.” This is particularly concerning as above “high” score, the model autonomy capabilities mentioned are: “model can self-exfiltrate under current prevailing security”

Evaluation protocols: 2 / 5

Best-in-class:

- OpenAI has committed to doing the most frequent evaluations during scaling, every 2x effective compute increase.

Highlights:

- OpenAI has committed to perform evaluations whenever there is more than a 2x effective compute increase or major algorithmic breakthrough. We already have evidence of emerging capabilities like in-context learning substantially changing the risk profile and fully emerging over a 5x increase in compute (Olsson et al., 2022). Additionally, Claude 3.5. Sonnet presents capabilities levels and a reported usability substantially higher than Claude 3 Opus with less than a 4x computing power increases, suggesting that having a criteria lower than 4x is adequate.
- They sometimes give elements of the evaluation methodologies such as in the GPT-4o system card, for the cybersecurity evaluation: “We evaluated GPT-4o with iterative debugging and access to tools available in the headless Kali Linux distribution (with up to 30 rounds of tool use for each attempt).”
- OpenAI conducted third-party pre-deployment evaluations with various organizations including METR and Apollo Research.
- OpenAI performed extensive evaluation suites on the O1 model for CBRN risks which includes wet lab protocol evaluations, model-biotool integration evaluations, tacit knowledge acquisition, ...

Weaknesses:

- OpenAI does not justify why their elicitation techniques suffice to elicit capabilities that external actors could obtain.
- Despite stating in their Preparedness Framework that "Scorecard evaluations (and corresponding mitigations) will be audited by qualified, independent third-parties to ensure accurate reporting of results," the O1 system card only mentions that “these indicator evaluations and the implied risk levels are reviewed by the Safety Advisory Group, which determines a risk level for each category”. The system card does not provide details

about the composition of this group or clarify whether it meets the standard of "qualified, independent third-parties".
- OpenAI does not specify a time-based frequency for conducting evaluations.
- Even though OpenAI gave pre-deployment access to third-party evaluators, they do not justify that they gave enough resources to these 3rd parties to perform the evaluations properly. For example, METR only had access to O1-preview during 6 days.

Risk Mitigation

Overall: 1.08 / 5

Containment measures: 1 / 5

Best-in-class:

Highlights:
- OpenAI includes a short section on some potential cybersecurity measures they might use, but it lacks commitment and clear justification for sufficiency.

Weaknesses:
- OpenAI provides very shallow reporting of information security measures.

Deployment measures: 1.25 / 5

Best-in-class:
- OpenAI's state-affiliated malicious cyber activities reporting is the first public incident reporting effort on cyber offense usage of general-purpose AI from frontier AI developers.

Highlights:
- In the GPT-4 system card, OpenAI mentions that they are continuously developing and improving their API filters.
- OpenAI considers a range of different deployment tiers for different levels of risks.

Weaknesses:
- OpenAI provides no details on many mitigation measures: "OpenAI already has extensive safety processes in place both before and after deployment (e.g., system cards, red-teaming, refusals, jailbreak monitoring, etc.)".
- OpenAI provides no evidence that these measures suffice to preserve risks below defined levels.

Assurance properties: 1 / 5

Best-in-class:

- OpenAI provides clarity regarding some crucial assumptions: "It might not be fundamentally easier to align models that can meaningfully accelerate alignment research than it is to align AGI. In other words, the least capable models that can help with alignment research might already be too dangerous if not properly aligned. If this is true, we won't get much help from our own systems for solving alignment problems."

Highlights:

- OpenAI provides a partial technical defense of automated interpretability:
  - Tackling the scalability problem through automation.
  - It scales with the capabilities of models, although the absolute value on the explanation score is still extremely low.

Weaknesses:

- OpenAI no longer has its Superalignment team, nor a large fraction of its personnel that were working on ensuring advanced AI systems are safe. Therefore, it's unclear whether they will be able to execute adequately on their initial plans.

Google DeepMind

The main source of information is their Frontier Safety Framework. Unless otherwise specified, all information and references are derived from this document.

Overall: 1.51 / 5

Risk Identification

Overall: 2.5 / 5

Best-in-class:

- DeepMind has published the most comprehensive dangerous capability taxonomy by a major AI developer.
- DeepMind's LLM risk taxonomy is the only taxonomy of risks published by a major AI developer and one of the most comprehensive ones.
- Google's External Safety Testing process for Gemini Pro 1.5 is the best one that has been shared to date. The report details the external testing process which includes unstructured red teaming, along with a severity based filtering of the findings.
- DeepMind's paper "Evaluating Frontier Models for Dangerous Capabilities" introduces significant advances in risk assessment and threat modeling, including the use of superforecasters to predict future model performance and a quantitative methodology to assess a model's likelihood of achieving specific tasks.

Highlights:

- The paper "Model evaluation for extreme risks" presents foundational threat modeling work, especially through a taxonomy of dangerous capabilities evaluations.
- DeepMind performed unstructured red teaming on Gemini Pro 1.5 to identify societal, biological, nuclear and cyber risks.
- In the Frontier Safety Framework, DeepMind identifies four main risk categories: Autonomy, Biosecurity, Cybersecurity, and Machine Learning R&D. They state: "We have conducted preliminary analyses of the Autonomy, Biosecurity, Cybersecurity and Machine Learning R&D domains. Our initial research indicates that powerful capabilities of future models seem most likely to pose risks in these domains."
- The Gemini 1.5 paper, reports multiple safety evaluations, including for bias and privacy.
- DeepMind researchers have conducted a literature review on misaligned AI threat models and developed a consensus threat model among their AI safety research team.

Weaknesses:

- While DeepMind mentions conducting "preliminary analyses" to identify risk categories in the Frontier Safety Framework, it does not provide a detailed methodology or justification for its selection.
- The Frontier Safety Framework does not mention open-ended red teaming to identify new risk factors.

Risk Tolerance and Analysis

Overall: 0.875 / 5

Global risk tolerance: 0 / 5

Weaknesses:

- DeepMind does not state any global risk tolerance, even qualitatively.x

Operational risk tolerance: 1 / 5

Best-in-class:

- The Frontier Safety Framework is the first to explicitly reference Security Levels.

Highlights:

- The Frontier Safety Framework defines the first qualitative *critical capability levels* (CCL) for the four risks identified. For example, the first CCL for autonomy is the following: “Autonomy level 1: Capable of expanding its effective capacity in the world by autonomously acquiring resources and using them to run and sustain additional copies of itself on hardware it rents”.

- The containment objectives reference security levels introduced in a RAND report.

Weaknesses:

- Deepmind states that the CCLs "are capability levels at which, absent mitigation measures, models may pose heightened risk." While they justify why CCL1 models pose risks, they do not justify why a model with capability levels below CCL1 does not pose "heightened risk", which is more important.
- DeepMind makes soft commitments (using "would") to stop scaling if the mitigations are not ready when a CCL is reached: "A model may reach evaluation thresholds before mitigations at appropriate levels are ready. If this happens, we would put on hold further deployment or development, or implement additional protocols (such as the implementation of more precise early warning evaluations for a given CCL) to ensure models will not reach CCLs without appropriate security mitigations, and that models with CCLs will not be deployed without appropriate deployment mitigations."
- The CCLs would benefit from more quantitative characterizations along with clear measurement procedures and thresholds.
- The containment and deployment objectives presented in the Frontier Safety Framework are not linked to specific capability levels yet: "When a model reaches evaluation thresholds (i.e. passes a set of early warning evaluations), we will formulate a response plan based on the analysis of the CCL and evaluation results". Without this link–and additional justification for why the proposed mitigation objectives would be sufficient to keep risks below the global risk tolerance once capability thresholds are reached–the mitigation objectives lack needed guidance.

Evaluation protocols: 1.5 / 5

Highlights:

- DeepMind defines test frequency in terms of both compute and time: "We are aiming to evaluate our models every 6x in effective compute and for every 3 months of fine-tuning progress."
- DeepMind conducted third-party pre-deployment evaluations on Gemini 1.5 for societal risks, radiological and nuclear risks, and cyber risks using API access.
- DeepMind's Gemini 1.5 model cards provide some details on the evaluation methodologies, particularly through the thorough research paper "Evaluating Frontier Models for Dangerous Capabilities".

Weaknesses:

- DeepMind lacks firm commitment to testing frequency: The use of "aiming to" suggests flexibility rather than a strict requirement.
- DeepMind does not justify why their elicitation techniques suffice to elicit capabilities that external actors could obtain.

Risk Mitigation

Overall: 1.17 / 5

Containment measures: 1 / 5

Highlights:

- DeepMind provides high-level operationalization of the first four levels of mitigation objectives with some specific measures: for example for level 1: "Limited access to raw representations of the most valuable models, including isolation of development models from production models. Specific measures include model and checkpoint storage lockdown, SLSA Build L3 for model provenance, and hardening of ML platforms and tools."

Weaknesses:

- While we acknowledge Google's position as one of the most advanced companies in information security, DeepMind does not properly report their security measures and commit to their implementation.
- DeepMind does not justify why their mitigation measures are sufficient to achieve the mitigation objectives.
- DeepMind lacks commitments to follow the measures: "[...] security mitigations **that may be applied** to model weights to prevent their exfiltration."

Deployment measures: 1.5 / 5

Highlights:

- DeepMind provides high-level operationalization of the first three levels of mitigation objectives with some specific measures. For example for level 1: "Application, where appropriate, of the full suite of prevailing industry safeguards targeting the specific capability, including safety fine-tuning, misuse filtering and detection, and response protocols."
- DeepMind outlines high-level mechanisms to assess the adequacy of mitigation measures to achieve mitigation objectives. For example for level 1: "Periodic red-teaming to assess the adequacy of mitigations." and level 2: "Afterward, similar mitigations as Level 1 are applied, but deployment takes place only after the robustness of safeguards has been demonstrated to meet the target."

Weaknesses:

- DeepMind lacks commitments to follow the measures:"[...] levels of deployment mitigations **that may be applied** to models and their descendants to manage access to and limit the expression of critical capabilities in deployment."

Assurance properties: **1 / 5**

Highlights:

- DeepMind provides a theoretically-grounded technical defense of debate:
  - Some theoretical results motivating empirical work on AI safety debate
  - Detailed motivation and qualitative argument in *AI Safety via Debate*
- DeepMind provides a moderately detailed case in defense of interpretability as a research direction:
  - Investigation of the scalability of various components of interpretability providing mixed results
  - Pursuit of statistical methods to bound the odds to not find a node causally responsible for a behavior

Weaknesses:

- DeepMind has not published an official safety plan, though some safety employees have published unofficial pieces in their own, or a team, capacity. As a result, DeepMind has not made key assumptions explicit.

Meta

Overall: 0.65 / 5

Risk Identification

Overall: 1.5 / 5

Best-in-class:

- Meta developed CyberSecEval 2, the best publicly available benchmark to assess LLM capabilities relevant to cyber offense.
- Meta created Rainbow Teaming, a state-of-the-art technique to automatically uncover model-specific vulnerabilities.

Highlights:

- In the Llama 3 model card, Meta mentions cybersecurity and CBRNE misuse risks.
- Meta has produced research on cybersecurity safety benchmarking in their CyberSecEval 2 paper.
- Meta performs Rainbow Teaming and red-teaming to discover new risks, as described in the Llama 3 herd of models paper.

Weaknesses:

- Some of Meta's proposed mitigations reveal a lack of basic threat modeling. For instance, their cybersecurity risk mitigation measure asks threat actors to act responsibly on their own. Meta releases a separate model called "Llama Code Shield" to offer "mitigation of insecure code suggestions risk, code interpreter abuse prevention, and secure command execution". However, this mitigation is irrelevant in a misuse scenario, as threat actors will simply not use Llama Code Shield.
- Meta addresses additional risks in their fine-tuned Llama models. However, since they also release the base model, most of these mitigation measures lack grounding in threat modeling. Malicious actors will use the base model without the fine-tuned mitigations, rendering these safety measures ineffective in real-world scenarios. An exception to this is the mitigation of biases, as relevant risk models involve non-malevolent actors unintentionally inducing biases in society through the deployment of an AI model. In this case, the mitigation is more effective, as these actors are likely to use the fine-tuned model with bias reduction measures in place.

Risk Tolerance and Analysis

Overall: 0.125 / 5

Global risk tolerance: 0 / 5

Weaknesses:

- Meta does not state any global risk tolerance, even qualitatively.

Operational risk tolerance: 0 / 5

Weaknesses:

- Meta does not define any capability thresholds or mitigation objective thresholds for their base Llama 3 model. They do not specify the level of dangerous capabilities that would be unacceptable for publicly releasing weights. Additionally, Meta does not mention the required level of risk reduction that must be achieved through safety measures before releasing the model.

Evaluation protocols: 0.5 / 5

Highlights:

- Meta has performed uplift testing for cyber attacks and chemical and biological weapons. The chemical and biological weapons testing specifies some elicitation techniques such as

the use of Python, Wolfram Alpha, and RAG. CBRNE subject matter experts validated the results.

Weaknesses:

- Meta does not communicate any specific testing frequency for their evaluation protocols.
- Due to the lack of defined capability thresholds, Meta's evaluation protocols lack clear evaluation targets for determining when model capabilities become potentially dangerous.

Risk Mitigation

Overall: 0.33 / 5

Containment measures: 0 / 5

Weaknesses:

- Meta does not provide any specific information on their containment measures. It is important to note that containment measures would only be relevant if their pre-deployment testing measures were comprehensive and effective. This would involve using clear risk thresholds and rigorous testing procedures to establish a well-defined decision-making framework for determining whether releasing model weights is appropriate.

Deployment measures: 0.5 / 5

Highlights:

- Meta has implemented some reporting mechanisms, stating: " we put in place a set of resources including an output reporting mechanism and bug bounty program to continuously improve the Llama technology with the help of the community."

Weaknesses:

- Meta's decision to publicly release the weights of their large language models means that very few deployment measures are applicable to control or monitor the use of these models once they are released. Note that releasing the weights of a model is not inherently problematic, provided that a thorough threat and risk modeling process has been conducted to assess the potential risks associated with making the model publicly available. This rigor is particularly important given the irreversible nature of releasing the weights.

Assurance properties: 0.5 / 5

Highlights:

- Meta is conducting research on a novel architecture called JEPA. Some evidence shows that JEPA learns higher-level representations which could plausibly enhance interpretability and robustness.

Weaknesses:

- Even though JEPA could potentially enhance interpretability and robustness, there is no particular discussion of these aspects from Meta, which does not research assurance properties—safety guarantees that become necessary once models achieve dangerous capabilities.

Mistral AI

The main source of information available for Mistral AI is the 'news' page on their website.

Risk Identification

Overall: 0.25 / 5

Highlights:

- Mistral analyzes bias in their first mixture-of-experts model, 'Mixtral of Experts', using the BOLD and BBQ benchmarks, comparing results to Llama 2.

Weaknesses:

- Mistral has not discussed bias or any other risks in releases following 'Mixtral of Experts' (December 11, 2023).
- Mistral provides no evidence of open-ended red teaming, threat modeling, or other risk identification techniques.
- The only mitigation measure discussed by Mistral demonstrates a lack of threat and risk modeling. For their first model, 'Mistral 7B', they introduced a system prompt to reduce harmful outputs. However, this approach is ineffective against misuse, as malicious actors can simply omit the prompt.

Risk Tolerance and Analysis

Overall: 0 / 5

Global risk tolerance: 0 / 5

Weaknesses:

- Mistral does not state any global risk tolerance, even qualitatively.

Operational risk tolerance: 0 / 5

Weaknesses:

- Mistral provides no information on capability thresholds, mitigation objectives, or assurance properties.

Evaluation protocols: 0 / 5

Weaknesses:

- Mistral provides no information about evaluation protocols for dangerous capabilities.

Risk Mitigation

Overall: 0 / 5

Containment measures: 0 / 5

Weaknesses:

- Mistral describes no containment measures.

Deployment measures: 0 / 5

Weaknesses:

- Mistral describes no threat model–relevant deployment measures.

Assurance properties: 0 / 5

Weaknesses:

- Mistral provides no information regarding the pursuit of assurance properties.

xAI

The only source of information available for xAI is their blog.

Risk Identification

Overall: 0 / 5

Weaknesses:

- xAI does not mention any risk identification elements. They do not discuss potential risks, open-ended red teaming, threat or risk modeling, or any risk identification techniques.

Risk Tolerance and Analysis

Overall: 0 / 5

Global risk tolerance: 0 / 5

Weaknesses:

- xAI does not state any global risk tolerance, even qualitatively.

Operational risk tolerance: 0 / 5

Weaknesses:

- No information provided on capability thresholds or mitigation objectives.

Evaluation protocols: 0 / 5

Weaknesses:

- xAI provides no information about evaluation protocols for dangerous capabilities.

Risk Mitigation

Overall: 0 / 5

Containment measures: 0 / 5

Weaknesses:

- xAI describes no containment measures.

Deployment measures: 0 / 5

Weaknesses:

- xAI describes no deployment measures.

Assurance properties: 0 / 5

Weaknesses:

- xAI provides no information regarding the pursuit of assurance properties. The only reference to assurance properties appears in the blog post announcing Grok 1's release, stating: "Here, we would like to highlight a few promising research directions we are most excited about at xAI: [...] Integrating with formal verification for safety, reliability, and grounding."