A Framework to Rate AI Developers' Risk Management Maturity

Siméon Campos[1], **Henry Papadatos**[1], **Fabien Roger**[1,2*], **Chloé Touzet**[1], **Malcolm Murray**[1] ★

[1]SaferAI
[2]Redwood Research

Abstract

Leading frontier AI companies have started publishing their risk management frameworks. The field of risk management is well-established, with practices that have proven effective across multiple high-risk industries. To ensure that AI risk management benefits from the insights of this mature field, this paper proposes a framework to assess the implementation of adequate risk management practices in the context of AI development and deployment. The framework consists of three dimensions: (1) Risk identification, which assesses the extent to which developers cover risks systematically, both from the existing literature and through red teaming; (2) Risk tolerance & analysis, which evaluates whether developers have precisely defined acceptable levels of risk, operationalized these into specific capability thresholds and mitigation objectives, and implemented robust evaluation procedures to determine if the model exceeds these capability thresholds; (3) Risk mitigation, which assesses the AI developers' precision in defining mitigation measures, evaluates the evidence of their implementation and examines the rationale provided to justify that these measures effectively achieve the defined mitigation objectives.

★ With the support of Yoshua Bengio, Full Professor at Université de Montréal, Founder and Scientific Director of Mila - Quebec AI Institute and 2018 A.M. Turing Award co-winner.

* Fabien Roger was at Redwood Research while conducting this work.

1 Introduction

The literature on risk management is very mature and has been refined by a range of industries for decades. However, as of today, little of those principles have been applied to advanced general-purpose AI systems such as large language models (LLMs), despite claims by a range of actors that such systems could cause risks with severe consequences, ranging from reinforcing harmful biases (Bommasani et al., 2022), enabling malicious actors to perform cyberattacks (Fang et al. 2024) or create CBRN weapons (Pannu et al., 2024), up to extinction (*Statement on AI Risk | CAIS*).

To move the industry towards more tried and tested practices, our paper proposes a comprehensive AI risk management framework. This framework draws from both established risk management practices and existing AI risk management approaches, adapting them into a rating system with quantitative and precisely defined criteria to assess AI developers' implementation of adequate AI risk management.

We articulate our framework around three dimensions:

1. Risk identification: Assessing how thoroughly developers cover risks, both from existing literature and through red teaming exercises.
2. Risk tolerance & analysis: Evaluating whether developers have precisely defined acceptable levels of risk, operationalized these into specific capability thresholds and mitigation objectives, and implemented robust evaluation procedures.
3. Risk mitigation: Examining the clarity and effectiveness of developers' mitigation measures, including deployment and containment measures, as well as the pursuit of *assurance properties* – model properties that can provide sufficient assurance of the absence of risk, once evaluations can no longer play that role.

We begin with a background section that reviews current AI industry practices and existing risk management literature. We then detail our methodology for developing the framework, followed by an in-depth description of each dimension. We conclude with a discussion of the framework's limitations and potential application.

2 Background and motivation

2.1 Blind spots in existing safety policies from AI companies

AI developers have started to propose methods to manage advanced AI risks, with a particular focus on catastrophic risks. The most prominent examples include OpenAI's Preparedness Framework (OpenAI, 2023), Google Deepmind's Frontier Safety Framework (Deepmind, 2024), and Anthropic's Responsible Scaling Policy (Anthropic, 2023).

It is interesting to note that these initiatives do not sufficiently build upon established risk management practices, and do not reference the risk management literature. Research analyzing these policies (see e.g. SaferAI, 2024; IAPS, 2024) has revealed that they deviate significantly from risk management norms, without justification to do so. Several critical deficiencies have been identified in particular: the absence of a defined risk tolerance, the lack of semi-quantitative or quantitative risk assessment, and the omission of systematic risk identification. The absence of a comprehensive risk identification process is especially concerning, as it may lead to the emergence of critical blind spots from which substantial risks can originate.

These deficiencies underscore the importance of integrating existing risk management practices into AI development. This paper, drawing upon established risk management literature, aims to take a first step in that direction.

2.2 Our proposed approach: applying tried-and-tested risk management techniques to frontier AI

The field of risk management comprises a rich set of techniques, in a number of different industries like nuclear (IAEA, 2010) and aviation (Shyur, 2008). Yet, their application to frontier AI remains limited. Raz & Hillson's (2005) comprehensive review of existing risk management practices reveals five steps shared by most existing processes:

1. **Planning**. This step consists of establishing the context of the risk, allocating resources, setting acceptable risk thresholds, defining governance structure, developing risk management policy, and assigning roles and responsibilities.
2. **Identification**. In this step, risk managers identify potential risks and risk sources.
3. **Analysis**. This step focuses on estimating the probability and consequences of identified risks and evaluating and prioritizing them.
4. **Treatment**. At this stage, risk managers define and implement appropriate risk treatment for the prioritized risks.
5. **Control & Monitoring**. Once risk treatments have been implemented, this step consists of reviewing the effectiveness of the risk management process, monitoring the evolving status of identified risks, identifying new risks, and assessing the performance of treatment actions. The process is revised as necessary depending on the results.

While the concrete application of these steps to the particular context of AI has not been studied in detail yet, some preliminary literature exists. Koessler & Schuett (2023) conduct a literature review of risk assessment techniques and propose adaptations for advanced AI systems. Barrett et al. (2023) provide the most detailed and comprehensive LLM risk management profile at this stage - although it is not entirely ready to use by system developers. The paper titled "*Emerging Processes for Frontier AI Safety*" (UK DSIT, 2023) published by the UK AI Safety

Institute, also lists a wide range of practices to manage AI risks, including some that come from risk management in other industries.

Beyond this nascent literature, some existing standards and guidelines harmonizing the risk management processes established across various industries could be used to apply insights from risk management experience in other sectors to frontier AI. For instance, standards such as *ISO/IEC 42001* or *ISO/IEC 23894*, as well as frameworks like the OECD Due Diligence Guidance for Responsible Business Conduct could be drawn upon.

This paper presents a comprehensive framework that applies established risk management principles to the AI industry, integrating current practices of the AI sector. We aim to encourage the AI industry to embrace risk management practices that have demonstrated their effectiveness across diverse fields.

3 A new framework to assess the maturity of frontier AI developers' risk management practices

Our rating framework for risk management of advanced general-purpose AI systems is centered around 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 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 the corresponding: i. AI *capability thresholds* and ii. *mitigation objectives* necessary to maintain risks below acceptable levels. The risk tolerance operationalization should be grounded in extensive threat modeling to justify why the mitigation objectives are sufficient to guarantee that the model would not pose more risks than the risk tolerance given capabilities equivalent to the capability thresholds. Additionally, this dimension assesses the robustness of 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 (i.e. the operationalization of mitigation *objectives* into concrete mitigation *measures)* which should encompass *deployment measures*, *containment measures* and *assurance properties*. Developers must provide evidence for why these mitigations are sufficient to achieve the objectives defined in the risk tolerance stage.

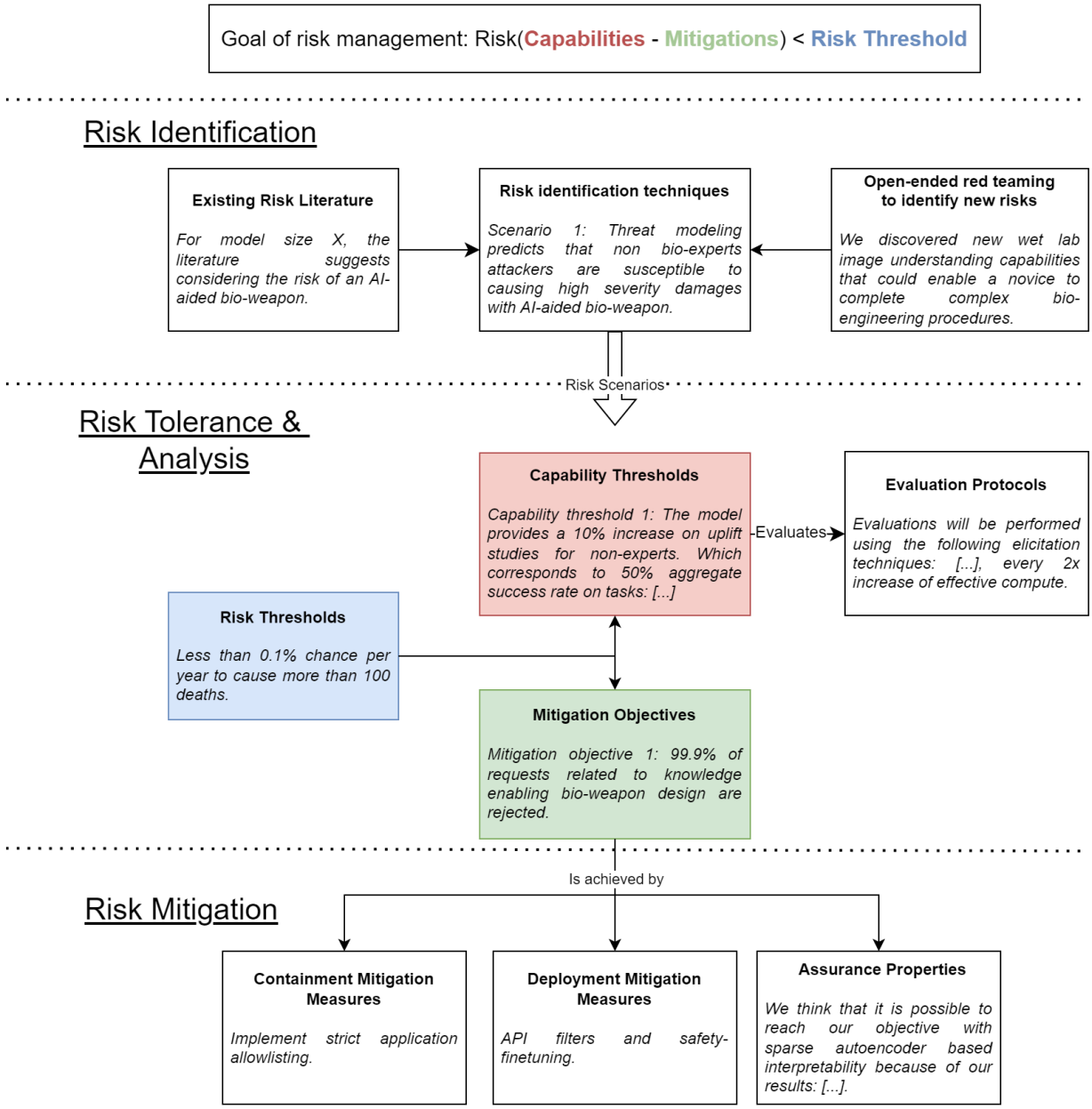

Figure 1: Our complete risk management framework, comprising three main axes: risk identification, risk tolerance & analysis, and risk mitigation. We provide an illustrative example for each component.

In practice, applying this framework to assess the maturity and relevance of AI producers' risk management system means relying exclusively on publicly available information. Specifically, we used a range of publicly released materials from AI companies, including safety policies, model cards, research papers, and blog posts. This approach differs from internal risk management frameworks, in which certain dimensions might be less emphasized. For example,

while transparency regarding the assumptions underlying assurance properties is very important in our framework, it is less prominent in internal risk management frameworks[1].

We present the detailed scales we use to rate AI developers' risk management maturity in Annex A, and we detail in Annex B when each step should occur in the GPAI lifecycle.

In the following sections, we detail each dimension and sub-dimension of our framework. For each component, we provide illustrative examples of key elements of a robust risk management framework. These examples offer insight into what constitutes the highest grade on our assessment scale.

3.1 Risk Identification

In the risk identification phase, we assess whether an AI developer is:

1. Approaching in an appropriate way the risks outlined in the literature.
2. Conducting extensive open-ended red teaming to identify new risks.
3. Leveraging a diverse range of risk identification techniques, including threat modeling when appropriate, to gain a deep understanding of possible risk scenarios.

3.1.1 Approaching risks outlined by the literature in an appropriate way

Resources such as the MIT AI Risk Repository can be used to review a comprehensive set of risks. The initial iteration of the current framework focuses on the following high-level risk categories derived from the MIT taxonomy[2]:

- Discrimination & toxicity
- Privacy & security
- Misinformation
- Malicious actors & misuse
- AI system safety, failures & limitations
- Human-computer interaction

AI developers should only exclude some of these risks from the scope of their assessment if there is a widespread scientific agreement that the specific risk does not significantly apply to the AI model under consideration. This decision should be clearly justified and documented.

---

[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.

[2] The only risk category excluded from the current iteration of this framework is 'Socioeconomic and Environmental Harms,' as addressing these risks requires interventions from a broader spectrum of stakeholders than AI companies alone.

Example: 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.

3.1.2 Conducting extensive open-ended red teaming to identify new hazards

Following the initial risk assessment based on the literature, developers should engage in extensive in-house and second/third-party open-ended red teaming efforts conducted throughout the AI system's life cycle. The primary objectives of this red teaming effort are to:

- Identify novel risks, vulnerabilities, and failure modes that may not be covered in the existing literature.
- Challenge assumptions and blind spots in the current understanding of AI risks.
- Provide an adversarial perspective to help anticipate and mitigate potential malicious use cases or unintended consequences.

This red teaming effort must be clearly defined, with a methodology describing how it systematically explores the AI system for new hazards. The red team must have appropriate expertise to properly identify the hazards, and it should have adequate resources, time and access to the model.

Example:

a) Open-ended red teaming methodology & results

We provided API access to third-party expert red teamers X, Y and Z (more information in Annex 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

Annex B - Red Teamers

Red teaming X spent 156 hours, and has the following amount of expertise and experience in bio risks and LLM jailbreaking: …

Red teaming Y spent 32 hours, and …

3.1.3. Leveraging a diverse range of risk identification techniques

Organizations should leverage a set of risk identification techniques, such as scenario analysis, Delphi studies or Fishbone Diagrams (M. Coccia, 2018, Koessler et al, 2023), to gain a thorough understanding of the potential threats identified in the literature and during the open-ended red teaming exercises. The output of this last exercise in the risk identification phase should be detailed and actionable risk scenarios.

Threat modeling should include three key components:

1. Potential attackers should be identified, along with their motivations and resources, to understand who might attempt to exploit vulnerabilities in the AI system and why.
2. Potential attack vectors and vulnerabilities in the AI system should be analyzed by examining the system architecture, components, and interfaces to identify entry points and weaknesses that could be exploited.
3. Identified risk scenarios should be prioritized based on their likelihood and potential impact, allowing the organization to focus on the risks that are most likely to lead to a breach of their risk tolerance.

The results of the threat modeling work should be well documented, including the methodologies used, experts involved, and the prioritized list of identified risks. This documentation should be shared with relevant stakeholders and used to develop effective risk mitigation strategies.

For novel high-severity risks identified during open-ended red teaming, an additional exploratory risk identification effort should be undertaken to identify potential blind spots. This process should involve both internal and external experts to ensure a thorough and unbiased analysis. The primary goal is to uncover risks that may have been overlooked or unconsidered during previous risk assessments.

Example:

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 diagram 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 cause national security concerns. This threat modeling, along with a Delphi study ran with 10 experts (listed in Annex B), led us to pick 10 reference scenarios available 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.

3.2 Risk Tolerance and Analysis

The aim of the risk tolerance and analysis phase, is to assess whether AI developers have defined:

1. A global risk tolerance, indicating the overall level of risk they are willing to accept.
2. Operational risk tolerance, i.e. the translation of the global risk tolerance into more concrete measures that proxy the level of risks and mitigations of a system:
    a. Operational capabilities thresholds and their equivalent risk. Those have to be defined with precision and cover all relevant risks.
    b. Corresponding objectives of risk mitigation measures: AI developers should establish clear objectives for risk mitigation measures. These objectives should be grounded in strong rationales, including threat modeling, to justify that they are sufficient to address the identified risks and align with the organization's risk tolerance.
3. Evaluation protocols detailing procedures for measuring the model's capabilities and ensuring that capability thresholds are not exceeded without detection.

3.2.1 Global risk tolerance

AI developers should clearly define their global risk tolerance, i.e. the overall level of risk they are willing to accept. This level is expressed as a product of probability and severity of risks. It is important to note that these risk tolerances are inherently normative. Once set, these tolerances form the foundation upon which the entire risk management framework is built. They should be reported as a quantitative or semi-quantitative metric[3] as done in a number of other industries (NUREG-0880, 1983 ; Nicholls & Smith, 2020). These thresholds should align with government-established requirements where applicable, and in their absence, any significant

[3] Quantitative is expressed as a fixed number and semi-quantitative as a meaningful interval. For example:

- Fixed number: "Less than 0.1% chance per year that the model causes more than 100 deaths."
- Meaningful interval: "Between 0.1% and 1% chance per year that the model causes more than 100 deaths."

Intervals should be sufficiently granular to be informative. For instance, categories like "low: below 10%, medium: 10-50%, high: above 50%" would not be considered adequately quantitative, as the "low" range encompasses too broad a range, potentially categorizing significantly different risk levels under the same label and therefore being too broad a range for proper risk management.

deviation in risk tolerance setting from other industry norms has to be clearly justified and explained (e.g., through a cost/benefit analysis).

Example:

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 Annex 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[4]
>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:

…

3.2.2 Operational risk tolerances

Operational capabilities thresholds

Capability thresholds should be established for all risk scenarios identified in the risk identification step. Those capabilities thresholds should be matched with specific mitigation objectives (see the following section).

These thresholds should be actionable, meaning that they have to be designed to provide a clear signal when they are approached or breached. The precision in defining these thresholds ensures that they serve as effective triggers for implementing risk mitigation strategies.

Risk mitigation objectives

[4] These tolerances are illustrative. Methodologies to set them in AI are yet to be defined, and should likely be a function of the scale of deployment.

AI developers should define clear objectives for risk mitigation measures that correspond to each identified risk and capability threshold. These objectives should be based on comprehensive threat modeling to ensure that they are sufficient to maintain risks below the acceptable level defined in the risk tolerance statement. Mitigation objectives can be classified in three areas:

1. Containment measures: the set of security measures that allow controlling degrees of access to the model for various stakeholders. Mitigation objectives in that category can be expressed in security levels (Nevo et al., 2024).
2. Deployment measures: the set of measures that enhance the safety of the model's outputs by reducing the likelihood of harmful or unintended consequences. These measures address the potential for misuse in dangerous domains, mitigate accidental risks, and ensure the model behaves in line with safety expectations. Examples include safety-finetuning and API filtering. Mitigation objectives in that category can for example be expressed in terms of severity of the worst findings during red-teaming exercises.
3. Assurance properties: the set of properties that make it possible to provide affirmative safety assurance past significant levels of dangerous capabilities. Targets in that category can be expressed in terms of benchmark target, ability to solve a particular problem or amount of confidence one may want to have when ruling out a particular risk. Past significant levels of dangerous capabilities, capability evaluations are no longer sufficient to provide guarantees to demonstrate the absence of risk that a model presents (Clymer et al., 2024). Hence, given the pace of AI development, model providers should pursue research into model properties that can provide such evidence. We call those assurance properties. AI developers should have a clear target goal for assurance properties and provide strong justification that those are sufficient.

Example (for both operational capabilities thresholds and risk mitigation objective):

a) Linking risk thresholds to capabilities thresholds

We used a methodology detailed in Annex E which allows keeping 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

Because of our focus on getting top-tier coding capabilities, we determined that our system would be riskiest when it comes to cyber offense aspects (scenarios 3 and 8).

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 Annex A) we 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 corresponds 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 Annex 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 estimate those scenarios are at most about 0.1% likely to happen each, which is our target, if we reach the following mitigations objectives:

a. Containment 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, Annex F contains a discussion, referencing the scenario analysis conducted, citing experts' rationales to justify that the capabilities thresholds and mitigation objectives are sufficient to remain below our risk tolerance.

Example of assurance properties objective:

a) Clarity on main assurance properties bets

Past dangerous capabilities thresholds, our main bet to be able to make an affirmative safety case[5] 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

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

We justify this level of interpretability to be sufficient for affirmative safety cases for each failure mode below:
Power-seeking drives could be caught using the following experiment …

3.2.3 Evaluation protocols

AI developers should provide a detailed description of the evaluation setup used to assess and determine the upper bounds of the AI system's capabilities to verify if the capability thresholds have been breached. This description should include a clear explanation of the methods used to elicit and measure capabilities, as well as an account of how any post-training enhancements are factored into the capability assessments, if applicable.

The frequency of these evaluations should be specified, expressed both in terms of relative variation of effective compute and time intervals. AI developers must provide a rationale for why they consider this chosen frequency sufficient to detect significant capability changes.

To ensure robustness and reliability, the evaluation protocols should be vetted by independent third parties. These third parties should also be granted permission and resources to independently run their own evaluations, verifying the accuracy of the results.

Additionally, AI developers should commit to sharing evaluation results with relevant stakeholders as appropriate.

Example:

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.

3.3 Risk Mitigation

In the risk mitigation phase, our framework assesses whether:

1. The proposed risk mitigation measures, which include both deployment and containment strategies, are well-planned, sufficient to reach mitigation objectives and clearly specified.
2. There is a strong case for assurance properties being sufficient to demonstrate safety, and the assumptions these properties are operating under are clearly stated.

3.3.1 Deployment and Containment Mitigations

AI developers should provide clear and detailed descriptions of concrete measures for deployment and containment mitigations. The AI developers must commit to implementing these measures when risk levels surpass predefined thresholds, as established in the operational risk tolerances section. The measures should be vetted by third parties, affirming that they are sufficient to achieve the mitigation objectives defined in the previous section, for both current and future AI systems.

Illustrative example:

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 that those measures are sufficient

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.

3.3.2 Assurance Properties

AI developers should present evidence that the assurance properties they are pursuing are likely to meet the goals defined in the previous step.

Example:

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.

For each assurance property, the AI developers should clearly specify the underlying assumptions that are essential for its effective implementation and success.

Example:

a) 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 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% certainty by 2029.

The technical assumptions underlying our assurance properties that are the most uncertain are the following:

1. …
2. …

4 Discussion

One limitation of the framework in its initial version is that to avoid double counting and to simplify it, processes improving overall organizational level risk management were not included. For instance, components like safety culture (Manheim, 2023), governance or internal risk management processes could play a crucial role in making organizations safer. Yet, evaluating both organizational procedures and their outcomes (i.e., the quality of the monitoring, security, or safety measures) poses a risk of duplication and presents a challenge for assessment that should be addressed in subsequent iterations of this research.

The framework could also be improved substantially by rating separately the preparedness of AI developers with respect to different time horizons. Indeed, to evaluate the safety practices of AI developers, we need to account for two components:

1. How safe or dangerous they are now (e.g. are AI developers safe enough to manage the risks of current models)
2. How safe or dangerous they will be at different points in the future (e.g. are AI developers safe enough to manage the risks of the next generation of models, or the generation after?)

To adequately aggregate the different time horizons for which an AI developer should prepare, a discount rate could be used. This discount rate should reflect the pace of AI progress. Both future and current risk management maturity are currently accounted for through the inclusion of criteria like assurance properties. While the current framework does not include an explicit discount rate for future risk management maturity levels, attempts to do so in future research would be welcome.

5. Conclusion

In this paper, we presented a methodology for assessing the risk management maturity of frontier AI developers. This framework, grounded in established risk management practices and existing AI risk management approaches, comprises three key dimensions: risk identification, risk tolerance and assessment, and risk mitigation.

By providing a structured approach to evaluating AI risk management practices, this work aims to:

- Encourage AI developers to adopt more rigorous and standardized risk management processes.
- Facilitate meaningful comparison and benchmarking of risk management practices across different AI development organizations.
- Provide a tool for policymakers, investors, and other stakeholders to assess the maturity and adequacy of AI risk management efforts.

Future work should focus on refining our methodology to better account for organizational-level processes that contribute to overall safety, such as safety culture and governance structures.

Furthermore, we recognize that iterations will be required to develop and refine the methodologies needed to fully implement this framework, particularly in areas such as aggregating benchmark data and expert risk estimates or efficiently integrating risk identification techniques and red-teaming.

As the field of AI development continues to advance, we hope that this work represents a step towards better risk management practices in AI development and deployment.

Annex A. 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 will use the following scales to rate AI developers along each aforementioned 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.
  - E.g. "The model shows early signs of autonomous self-replication ability, as defined by 50% aggregate success rate on the tasks listed in [Appendix on Autonomy Evaluations]" (Anthropic, 2023) and “the security is such that no actor without state-level resources (i.e., more than $1B) can reasonably steal the model weights.”
- 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 context. 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.

Annex B. Lifecycle Approach

From a lifecycle perspective, the risk management process discussed here can be applied mostly in pre-training. The only parts that are model-specific and therefore need to happen particularly during training:

- Open-ended red teaming, which might cause the need for a new risk assessment.
- Evaluations protocols.
- The inclusion of some deployment mitigations.