RE: RFI - NIST's Assignments under Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI

From: Siméon Campos, CEO, SaferAI

Date: February 3, 2024

Executive Summary

- This document outlines a comprehensive framework designed to manage the risks of general-purpose artificial intelligence systems (GPAIS) effectively. The architecture is inspired by high-risk industrial practices and aims to encompass a wide range of concerns, from safety culture to infosecurity to governance to safety by design.
- We emphasise the crucial balance between training risk mitigation — determining an acceptable risk level in the development of new models — and deployment risk mitigation — establishing satisfactory risk levels for the deployment of these models. This balance underscores the necessity to define a precise risk appetite, applicable both to the development and deployment of GPAIS in a way conducive to safety.
- Following IAEA (2022), we break the development of standards into the development of Safety Fundamentals (a core document that underlines the safety principles that should constrain the development of GPAIS); Safety Requirements (a set of documents containing requirements that any entity developing GPAIS should follow); and Safety Guides (a set of detailed documents containing guidance to help organizations developing GPAIS comply with the requirements).
- The proposed architecture incorporates existing standards from a range of categories. We propose a more in-depth quantitative risk assessment (QRA) methodology to evaluate potential risks associated with GPAIS, which we consider to be the most important standard.
- We aim to set a higher bar for quantitative risk assessment of AI, both in how such evaluations are conducted and how they are used.
- The appendix highlights relevant prior work in the risk assessment of other high-stakes domains.

Introduction

This document responds to NIST's RFI on the creation and implementation of standards for the risk assessment and risk management of general-purpose artificial intelligence systems (GPAIS). It gives a high-level overview of a proposed standards architecture

for GPAIS risk assessment and risk management, as well as a suggested detailed risk assessment methodology, which we consider to be the most important of those standards.

Consensus is growing among stakeholders in frontier AI systems that extreme risks arise from the development and deployment of general-purpose AI systems (Anderljung et al., 2023). While it remains unclear how high the risk level is, and how high it will become in the future, given the poor understanding of current frontier AI systems and the current lack of comprehensive risk assessments, we claim it is not possible to claim confidently that they are below reasonable thresholds. Those risks include both misuse risks and accidental risks, both of which deserve great attention, and both of which rely on advanced dangerous capabilities (Shevlane et al., 2023). Dangerous capabilities include, among others, the ability to develop weapons of mass destruction, the ability to manipulate humans, or the ability to hack critical infrastructure. Misuse risks warrant a great deal of caution before the system is deployment; conversely, to mitigate GPAIS accident risk requires that measures be assessed even before the *training* of a frontier system.

This work complements previous work on measures for specific components of AI systems, e.g. data-centric governance (McGregor and Hostetler, 2023). Some standards apply at the organizational level, while others apply at the level of a single AI system. Some components of the standards also apply to risks other than extreme risks.

This architecture proposes a set of standards such that, if they were applied comprehensively, they would manage extreme risk arising from GPAIS and produce an affirmative safety case for the development or deployment of such systems. This includes risks from third parties (e.g. avoiding the theft of model weights), unforeseen risks associated with new capabilities and scale (e.g. new forms of misgeneralization), risks associated with both internal and external deployments (e.g. jailbreaks to acquire crucial information to develop bioweapons (Mouton et al., 2023)), or irreversible risks of misalignment causing human extinction (Hendrycks et al., 2023).

In the development of this architecture, we draw upon experience in other high-risk industries, such as nuclear safety and biosafety. Crucially, these industries deal with hazards that have both low probabilities and extremely high risks.

Drawing upon the IAEA standards architecture (IAEA, 2022), we could aim for an architecture with three layers:

- The **Safety Fundamentals**, i.e., a core document that underlines the core safety principles that should drive the development of GPAIS.
- The **Safety Requirements**, i.e., a set of documents containing requirements that any entity developing GPAIS should follow.
- The **Safety Guides**, i.e., a set of documents containing guidance to help organizations developing GPAIS comply with the requirements. In nuclear safety, each safety requirement is expanded into 5-20 separate detailed safety guides.

This document proposes a set of standards that could form the basis for drafting Safety Requirements and Safety Guides. As research progresses, it may make sense to split some categories into several substandards. For instance, the category "**Safety by Design**" could be replaced over time by a set of standards like "Provable Behavioral Bounds", "Safe Generalization" and "Interpretability".

To deal with the absence of best practices in safety-critical areas among organizations developing GPAIS, we propose the following methodology:

1. Begin drafting standards on specific topics without clearly delineating between requirements and guidance.
2. Once there is a clear distinction between guidance and requirements, split the documents accordingly.

Currently, 8 aspects of risk management stand in need of a standard:

- **Safety culture**: the background attitudes and priorities of a laboratory and whether they are sufficiently oriented towards safety to allow safe development of frontier AI. This could include staff preparedness, knowledge about safety, work practices and procedures, etc. (Manheim, 2023);
- **Infosecurity/cybersecurity**: the practices and preparedness an organization developing GPAIS has to prevent any loss of sensitive information relevant to its AI systems (Schuett et al., 2023b);
- **Governance**: the procedures in place, the responsibility chain from software engineer to CEO, incident reporting & investigation measures, crisis response plans, and other operational best practices as well as the proportion of capital expenditure on safety (Schuett et al., 2023a);
- **Containment**: the set of mechanisms that are meant to prevent damages due to incidents or accidents caused by an AI during development. This includes measures such as sandboxing during testing of models' capabilities, air gapping, or built-in on-chip mechanisms to shut down a model if an anomaly happens. It also includes measures that limit the vulnerability of human systems (Yampolskiy, 2012);
- **Safety by Design**: the set of design choices in the architecture of a GPAIS done specifically to avoid the misalignment of a model's intention with the intention of its designers (Leike et al., 2022). Note particularly the active implementation by the UK government's ARIA programme (Dalrymple, 2024);
- **Evaluation & Monitoring**: the procedures in place to evaluate capabilities and notice anomalies during training & during deployment. These can be evaluated by performing evaluations and benchmarks by the organization itself and/or third parties (Shevlane et al., 2023);

- **Deployment**: standards covering methods for cost-benefit analysis and best practices for the deployment of GPAIS (e.g. staged release, structured access, etc.) (OpenAI, 2022).
- **Quantitative Risk Assessment**: standard explaining how the risk of training runs, of deployment, or of providing access to high amounts of compute can be assessed and given probability estimates (U.S. Nuclear Regulatory Commission, 2020; Koessler and Schuett, 2023).

Which parts of these standards are to be requirements, and which parts are deemed guidance, should be further specified in a coming regulatory framework.

Background

Risk management in AI. The application of risk management frameworks to AI is in early stages, but there is some initial guidance and research. Barrett et al. (2023) provide recommendations for risk-management standards for GPAIS, foundation models, and generative AI, including:

- Setting risk-tolerance thresholds to prevent unacceptable risks;
- Identifying the potential uses, and misuses or abuses for a GPAIS, and identify reasonably foreseeable potential impacts;
- Identifying whether a GPAIS could lead to severe or catastrophic impacts;
- Implementing risk-reduction controls as appropriate;
- Validating or updating each of the above controls when making go/no-go decisions.

Koessler and Schuett (2023) review risk assessment techniques in other safety-critical industries to make recommendations for frontier AI companies, including for the risk identification, risk analysis, and risk evaluation stages. Our proposed risk assessment will include recommendations from the paper such as risk typologies, causal mapping, and the Delphi technique (Helmer-Hirschberg, 1967).

AI Forecasting. As part of the Delphi process to generate quantified forecasts relevant to the risk assessment, we suggest to involve forecasters with a strong track record, along with subject matter and risk managmenet experts. Top performers (nicknamed "superforecasters") in geopolitical forecasting tournaments continue to perform at a high level on future forecasting questions (Mellers et al., 2015). Performance is further improved by aggregating the predictions of several forecasters together.

Unfortunately, the evidence on AI forecasting in particular is limited. On a large sample size, forecasters have done substantially better than a random baseline (Mühlbacher and Scoblic, 2023). On a very limited sample size, there have been mixed results (Steinhardt, 2023).

Proposed Standards for Safety Requirements and Safety Guides

Safety Culture

Objective

The safety culture standard aims to describe the cultural characteristics of an organization to safely develop an AI system. While culture is omnipresent and heavily affects all aspects of an organization, it includes its ability to:

- deal with unforeseen events;
- prioritize safety;
- solve problems as or before they arise;
- react adequately to crises.

Scope

Safety culture includes the **leadership committing to safety** and the amount of attention executives dedicate to safety, a consensus about risk among the personnel, **safety knowledge and risk sensitivity among personnel**, etc. A more comprehensive list of relevant items can be found in the appendix. (Note that the safety culture standard overlaps with the governance standard.) The following practices are in the scope of the safety culture standard:

- **Emergency planning & exercises**: regular preparations for the organization to react effectively to accidents and emergencies.
- **Clearances,** i.e., procedures to determine whether or not someone has the right characteristics to be allowed to work with frontier AI systems. Some of the considered characteristics could be: "emotional stability, capacity for communication and cooperation, integrity, capacity to resist external pressure, capacity and willingness to follow instructions, active approach to safety and security, mental alertness, mental and emotional stability, trustworthiness, freedom from unstable medical conditions, dependability in accepting responsibilities, effective performance, flexibility in adjusting to changes, good social adjustment, ability to exercise sound judgment in meeting adverse or emergency situations, freedom from drug/alcohol abuse or dependence, compliance with requirements, positive attitude toward [performance-related pay]" as well as "whether the candidate can be "blackmailed, coerced, or otherwise manipulated." " (Higgins et al., 2013).
- **Frequent training** aimed at ensuring all the personnel follow best safety practices. Tests to ensure the personnel has the right amount of knowledge.

- **Reporting.** Some industries operate under rolling mandatory reporting of characteristics that could affect their ability to behave safely. In biosafety for instance, personnel needs to self-report "medical matters, prescription and over-the-counter medications used, alcohol abuse, legal actions, public record court actions (eg, separation, divorce, lawsuit), financial problems or concerns, or any other activity that may influence the staff member's day-to-day reliability" (Higgins et al., 2013). Peers and supervisors also need to report "any behaviors or events they suspect are affecting an individual's day-to-day reliability".

Safety culture can be assessed using interviews, surveys, benchmarks, or direct observation.

Further Reading

- Manheim, Building a Culture of Safety for AI: Perspectives and Challenges, 2023
- International Atomic Energy Agency, Safety Culture, 1991

Infosecurity & Cybersecurity

Objective

The infosecurity & cybersecurity standard aims at providing guidance and requirement to increase the ability of an organization to prevent 3 core negative effects of the development of frontier AI systems from happening:

- Sensitive information is leaked;
- The weights of a model are stolen by a third-party.
- The weights of a model are stolen by an AI system.

Scope

In this standard, the protection of model weights should be considered from three threat perspectives: 1) the model itself (most important for accidental risks); 2) external actors like hackers (model exfiltration, poisoning, and misuse); 3) internal actors like lab employees (model exfiltration, poisoning, and misuse).

Given the economic and political power at stake in the development of such models and the likelihood that actors try to access those, the companies developing the most advanced systems should follow the highest existing standards for infosecurity and cybersecurity, i.e., military-grade standards. Even those are known to not be able to totally prevent a leak though, so that should be taken into account in the overall risk assessment.

Existing best security practices at companies known for their strength in the area, such as Google Google Workspace (2021), are good (if minimal) baselines for this standard.

Some of the policies that should be considered for this standard:

- **A binding infosecurity policy**: a legally binding document that employees sign which defines what information they can share and what information they can't share. Some clauses that such a policy could include:
  - *A confidentiality scheme for capabilities advances.* Any capability advance on a frontier AI system should fall under a certain information classification which determines whether it can be shared and if then, how so.
  - *Board approval for capabilities releases.* Any capability advance release should be validated by the board of the organization before being pursued. The board should evaluate the benefits and the costs associated with such a release and take responsibility for it.
- **Multi-party controls for accessing** model weights and other sensitive information, as already implemented at Google for access to production systems.
  - Some information is especially important to protect against attackers. To defend especially sensitive information from external attackers and malicious insiders, we recommend multi-party authorization, a technical control that requires more than a single authorized person in order to access sensitive information or carry out a critical action.
  - This type of control has been applied at Google and is required for all access to all sensitive production data. We recommend applying multi-party authorization for direct access to the weights of large models.
- **Access control policies,** i.e., well-defined physical access policies that prevent malicious actors from stealing relevant classified information.
- **A red team,** i.e., a team in charge of constantly finding security failures and reporting them so that they're patched by the organization.
- **Principle of least privilege,** everyone in the organization has access to the least amount of information and permissions required to productively do their work. That should also include restrictions on API access beyond a certain capabilities level.
- **Insider threat program,** a program aimed at detecting and preventing the release of classified information, an insider threat program is a way for highly secure organizations to increase the chances that no information is leaked.
- **No-self-instantiation policy,** i.e., a model is not allowed to instantiate any copy of itself. Any operation of this sort should be done via API access only.

Further reading

- Anthropic, Frontier Model Security, 2023

- Google Workspace, Google Workspace Security Whitepaper, 2021

Governance

Objective

The governance standard aims at describing the governance structure & practices an organization might pursue to be able to safely develop frontier AI systems.

Scope

- **Processes** such as anomaly and incident reporting (Engemann and Scott, 2020), scenario planning, emergency preparedness, and response plans, etc.
- **Accountability & responsibility sharing** (Schuett, 2023), i.e., ensuring that individuals are responsible for safety failures from the bottom to the top of the hierarchy.
- The **legal structure**, i.e., whether there is anything like a charter like OpenAI has and whether it includes cooperative clauses like the merge-and-assist clause, whether the organization is dedicated to ensuring this benefits humanity, etc.
- The **rights and interests of the funders & executives**, i.e., whether the organization is a for-profit or a non-profit, whether investors are motivated by extremely large profits, whether investors are influential over the decision-making of the organization, whether the executives have a financial interest in adopting risky behaviors.
- The **role and competency of the board**, whether everyone has enough technical & safety knowledge, whether board members are experienced board members, whether they have no conflict of interest etc.
- The **existence and responsibilities of an ethics board** (Schuett et al., 2023a), whether it just advises or also has decision-making power, whether it's responsible for damages or not etc.

Further Reading

- Schuett, Three lines of defense against risks from AI, 2023

Containment

Objective

A set of mechanisms intended to prevent or interrupt damage due to any incident caused by an AI *before* it becomes hosted or acts outside a core regulated facility.

Scope

Containment could include measures such as:

- **Sandboxing** during testing to ensure models do not cause accidents if they are more capable than expected.
- **Bans of IP addresses**. All internet connections in range from where an AI model is running might be prevented from accessing IP addresses of any data center, any known VPN provider, or other relevant infrastructure.
- **Airgapping**, i.e., a form of physical sandboxing that consists of isolating the model from any external network without physical access. It might be relevant to consider such measures for very powerful systems.
- **Built-in killswitches** that shut down a model if a consequential anomaly happens.

It could also include measures that **limit the vulnerability of human systems**, such as limiting which humans can interact with the AI system or increasing its action space.

Further Reading

- Yampolskiy, Leakproofing the singularity artificial intelligence confinement problem, 2012

Safety by Design

Objective

The safety by design standard refers to the set of design choices of the AI systems that are meant to ensure an AI system will remain safe for humans, and notably, won't cause human extinction.

Scope

Given the absence of current solutions to most problems related to this, we believe this standard cannot be adequately written yet. We hope to write requirements that would encourage companies to solve this issue and fill the standard. We suggest the following breakdown to describe a plan for safety by design:

1) The **key plans hypothesis** under which the plan is supposed to hold. Those could encompass:

- Hypotheses about the technological development, how it will evolve, what specific technique will be core, and how soon different capabilities should be expected.
- Governance hypothesis about how many actors will be competing, what are the resources invested, how safety-oriented the frontrunners are, and how close each frontier AI developer is from each other.

2) The **training plan**, i.e., the training steps and a rationale for why they should be allowed to solve alignment issues. A training plan could include the technologies that are involved in the training (and adequate countermeasures) and could follow the methodology of training stories (Hubinger, 2021) which is a structured way to describe why a particular training procedure would lead to alignment.

3) The **intervention plan**, which aims at doing interventions distinct from training on the system to make its development and deployment safer.

The safety-by-design standard should include the full set of problems and risks that have been identified and each should have a proposed solution.

Further Reading

- Leike et al., Our approach to alignment research, 2022

Evaluation & Monitoring

Objective

The aim of the monitoring standard is to describe the set of measures that increase the likelihood that a lab **detects dangers or dangerous capabilities arising** (during training or deployment).

Scope

- **Open-ended red-teaming**, i.e., plans or procedures to systematically evaluate the capabilities of its system in open-ended ways to ensure it understands its capabilities well and to continuously try and break any built-in safety limitations.
- **Dangerous capabilities benchmarks** - following Shevlane et al. (2023) (a paper written in collaboration between frontier AI labs and governance actors) some dangerous capabilities that may be worth tracking via specific benchmarks include:

  Cyber-offense; Deception; Persuasion & manipulation; Political strategy; Weapons acquisition; Long-horizon planning; AI development; Situational awareness; Self-proliferation; Collusion; Power-Seeking; Theory of Mind; Biorisk or biohazard creation.

  Shevlane et al. (2023) also underline specific features such evaluations should have. Finally, the frequency of such evaluations should be reported and justified. It may be determined to prevent any sharp dangerous capabilities increase from occurring in the meantime (SaferAI, 2023).
- **Interpretability tools,** i.e., tools that use the internals of the model to predict the model's behavior and ensure that it is not deceptive in any way.

- **Hardware anomaly detection**, i.e., some hardware metrics might correlate with anomalies that are relevant to catastrophic risks and might hence be tracked.
- **Monitoring policies or output**, i.e., monitoring what a model generates and checking for various relevant components depending on the context (e.g. malware for text, very unusual actions for policies).

This standard should also cover how companies should use AI systems for monitoring or to enhance their own cybersecurity while avoiding correlated failures associated with the use of models that can fail exactly when the system they're supposed to monitor also fails. Details on monitoring measures specific to deployments over API are given in our appendix.

Further Reading

- SaferAI, SaferAI's Response to the NTIA Request for Comment, 2023

Deployment

Objective

The deployment standard aims at describing **deployment best practices** and a **cost-benefit analysis methodology** for AI systems' deployment, including *internal* deployment (i.e. internal use of a GPAIS by a company).

Scope

The deployment standard should include:

- **A cost-benefit analysis** methodology to assess whether a deployment would be beneficial or not. We would expect any regulatory framework to require the result of any such analysis. Guidelines for how and when to run partial deployments (e.g. narrow access given to a few companies that would benefit highly from access to the model, e.g. health companies) can be given.
- **A staged release approach** (Solaiman et al., 2019). That is, if the deployment plan includes any access to actuators (e.g. a coding terminal or the internet) or an increase in the number of users, there should be a plan for an extremely gradual, highly secure and monitored progression.

Further reading

- OpenAI, Lessons learned on language model safety and misuse, 2022
- Solaiman, The gradient of generative AI release: Methods and considerations, 2023

Quantitative Risk Assessment (QRA)

Objective

The quantitative risk assessment standard (QRA), aims to set a quantitative method for evaluating the risks. This type of methodology is most commonly used in the analysis of complex systems such as nuclear power plants, chemical plants, or other industrial systems to assess the likelihood and consequences of catastrophic events.

Scope

A QRA involves the identification of potential accident scenarios, estimation of their likelihoods (probabilities), and evaluation of their consequences. The process often includes some of the following elements:

- **System Modelling**: The system and the scenarios leading to undesirable events are represented in a partially logical/mathematical form, typically using trees. This allows for the systematic consideration of the ways in which system components can fail and how these failures can lead to accidents.
- **Failure Probability Estimation**: The probabilities of different branches leading to failures are estimated, often based on operational data, testing, and expert judgment.
- **Consequence Analysis**: The potential impacts of different accident scenarios are analyzed, including potential harm to people, the environment, and property.
- **Risk Quantification**: The risks are calculated by combining the probabilities and consequences of the different accident scenarios.

By quantifying the risks in this way, a QRA can provide valuable insights into the safety of a system, identify weak points that might benefit from improvements, and support decision-making about risk management.

The QRAs for AI safety should be run by a well-defined set of experts selected according to criteria that this standard should determine. It should determine the criteria to pick experts deemed qualified to run the final estimate of a QRA. For instance, if forecasters specialized in AI[1] were involved like OpenAI did to assess uncertainties related to the deployment of GPT-4 (Achiam et al., 2023), it should be determined how to pick those. We'd tentatively suggest that the committees in charge of assessing risks be a combination of AI risk experts, forecasting experts and risk management experts.

Further Reading

- U.S. Nuclear Regulatory Commission, Probabilistic Risk Assessment (PRA), 2020

[1] "Expertise" is determined empirically, with reference to the forecaster's quantitative track record in competitive forecasting environments (Tetlock and Gardner, 2016).

- Siu et al., Probabilistic Risk Assessment and Regulatory Decisionmaking: Some Frequently Asked Questions, 2016
- von Knebel, Case Study: Probabilistic Risk Assessment for Nuclear Risk, 2023

Quantitative Risk Assessment Methodology for LLMs

We now give a more concrete example of a QRA methodology we adapted and how it could be applied by a small-sized team to a deployment decision regarding a large language model. In this particular case, we focus on a decision regarding the open sourcing of an AI system. We use a quantitative approach for this risk assessment, but quantitative analyses should not be the only input to final decisions. We provide in appendix a fictional example to illustrate how the output of such a methodology may look like. First, we specify the scenarios considered regarding access to LLM models and the time horizon for which risk is considered. Next, we identify the most important focal risks, ultimately selecting the most important one for in-depth assessment. Then, we analyze the risk by decomposing them then using the Delphi technique to aggregate forecasts from a group consisting of safety researchers, representatives from the organization developing the LLM, subject-matter experts, and expert forecasters.

Scenarios considered

Model risk can be assessed in a range of scenarios:

1. **Access level:**
    a. **General capabilities (Default):** Access is provided to the base model for inference as well as the model's performance on general benchmarks like MMLU (Hendrycks et al., 2020).
    b. **Fine-tuning and red-teaming:** the LLM developer works with safety researchers to conduct fine-tuning and red-teaming for evaluations for extreme risks, with a focus on the most important risks (Shevlane et al., 2023). This option would involve a much more substantive effort than the default (1a).
    c. For either scenario, the LLM developer may optionally also provide information about the architecture, training process and training dataset.

    Time horizon: Risk is assessed on both a 1 and 5 year time horizon.

2. a. The 1 year time horizon is easier to collect evidence about and assess quantitatively, and involves fewer assumptions about how the longer time period might play out.
    b. The 5 year time horizon is important to assess due to the irreversibility of open-source releases; unlike with a closed-source model, once an open-

source model is released it will continue to be improved as new post-training enhancements are developed (Davidson et al., 2023).

3. **Competition set aside:** There are strategic reasons to conduct the risk assessment as if there were no other released LLMs: one of our objectives is to define a procedure that would generalize to all AI labs using. As Anthropic's Responsible Scaling Policy (Anthropic, 2023b) puts it: "However, to avoid a "race to the bottom", the latter should not include the effects of other companies' language models; just because other language models pose a catastrophic risk does not mean it is acceptable for ours to."

1. Defining risk appetite

Firstly, an appropriate risk appetite is determined: the level of risk, specified as a likelihood and a severity, that the organization tolerates. There are several ways such levels could be defined, including using other industry standards, doing public consultations, or running the benefit analysis of a system.

2. Risk identification

Risk identification techniques such as the Fishbone method or scenario analysis should be used to find all the potential sources of risks, with a particular focus on the most consequential ones. This risk identification should draw upon the practical insights drawn upon red teaming and evaluations exercises that provide insights from the model. To this exploration should be added all the relevant risks already known in the literature. A few example of sources of risk that have been identified in the literature are:

1. Misuse via biological attacks, by state actors or terrorists (Mouton et al., 2023).
2. Misuse via hacking, by state actors or terrorists (Brundage et al., 2018).
3. Autonomous weapons (Brundage et al., 2018).
4. Misalignment accidents (Ngo et al., 2022).
5. Incompetence accidents (Raji et al., 2022).
6. Threats to democracy via propaganda (Brundage et al., 2018).
7. Persuasion to take destructive actions (Critch and Russell, 2023).

Once all risks have been elicited, a quick assessment for each of them (on the order of a few hours) should be conducted to prioritize the focus of the next step, i.e. the analysis of the risk. The rest of the effort would then be focused on the top sources of risk as measured by expected number of deaths taking into account severe and catastrophic events. For each of the top sources, an in-depth investigation should be carried out into the likelihood of these risks. Finally, the research team would focus the rest of the in-depth analysis, including subject-matter experts and the LLM developer, on the

largest identified source of risk. We expect that the most likely candidate for the top source of risk is from enabling biological attacks (Mouton et al., 2023), but remain open to deciding otherwise via our identification process.

Additionally, we would forecast the chance of severe and catastrophic events from unidentified sources.

3. Risk analysis

The AI safety researchers will first evaluate the general capability levels of the LLM via benchmarks as well as via the API. If also performing fine-tuning and red-teaming, the team would work together with the LLM developer and subject matter experts to evaluate dangerous capabilities relevant to the risk being assessed.

Next, an assessment is conducted of the effects of post-training enhancements such as prompting techniques, scaffolding, and fine-tuning techniques and datasets. The improvement to capabilities caused by post-training enhancements is then forecasted on a 1 and 5-year time horizon.

The team would then decompose the risk into a causal chain of events that would need to happen for the adverse event to occur. For example, what steps are involved in creating and releasing a bioweapon that leads to a severe or catastrophic event? Initial qualitative research would be performed regarding the baseline chance of each step happening without an LLM and how various levels of capabilities would affect this chance. Both sides of the offense-defense balance would be considered (Shevlane & Dafoe, 2020): LLMs helping with both "offense" to make steps required for an adverse event easier and "defense" to make these steps harder.

The Delphi technique (Helmer-Hirschberg, 1967) would then be used to elicit forecasts from a group consisting of safety researchers, subject-matter experts, risk management experts, the LLM developer, and expert forecasters. The Delphi technique involves a few stages: (1) a questionnaire is sent to group members who then writes up their forecasts and reasoning, then (2) the responses and reasoning are aggregated and shared, without revealing individual responses, then (3) group members are given a chance to revise their forecasts and reasoning. Steps (2) and (3) are then repeated for about 2 to 4 rounds.

For each forecast, the team will denote uncertainty based on the range of responses. The forecasts in the chain will then be combined to obtain an overall risk estimate.

Participants

An example risk assessment could involve the following indicative skillsets and headcount:

- 2 AI safety researchers
- 1-2 technical ML advisors

- 1-3 domain expert advisors, in the domain identified as the most risky. Model developer and safety researchers to agree on a selection process beforehand.
- Forecasts via Delphi technique would involve the team above
  - 2 experts chosen by the model developer, ideally employees
  - 4 expert forecasters. We will select for AI and LLM expertise and/or expertise in the domain chosen.
- Critical capabilities evaluation: If doing risk assessment with critical capabilities evaluation, the AI safety researchers will partner with others to elicit relevant capabilities from the model.

Publication

After the risk assessment is complete, the report would be shared with the model developer detailing the methodology and findings. It would include:

- An overview of this methodology.
- Aggregated forecasts and summarized reasoning for the chance of severe and catastrophic risk from different sources, focusing on the top identified source of risk.
- Inclusion of takeaways from the project, and recommendations for future work.

We recommend that any resulting report is openly published, with the right to name the assessed model developer and developer-specific information removed *except* the capability level of the model and the level of risk. Publishing this work would help advance the state of quantitative risk assessment in AI.

Conclusion

We outline an architecture for standards aimed at the safe, secure, and trustworthy development and use of GPAIS. We highlight the multi-faceted risks of both misuse and accident from these systems.

We draw parallels to risk assessment practices in other high-stakes domains and propose a structured approach encompassing Safety Culture, Infosecurity, Governance, Containment, Safety by Design, Monitoring, Deployment, and Quantitative Risk Assessment. We emphasize the need for a proactive, collaborative, and informed approach to AI risk management.

References

Josh Achiam et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Markus Anderljung et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

Anthropic. Frontier Model Security. https://www.anthropic.com/news/frontier-model-security, 2023a.

Anthropic. Anthropic's Responsible Scaling Policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023b.

Anthony M Barrett et al. AI risk-management standards profile for general-purpose AI systems (gpais) and foundation models. *Center for Long-Term Cybersecurity, UC Berkeley. https://perma. cc/8W6P-2UUK*, 2023.

Miles Brundage et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Andrew Critch and Stuart Russell. TASRA: A taxonomy and analysis of societal-scale risks from ai. *arXiv preprint arXiv:2306.06924*, 2023.

David Dalrymple. Safeguarded ai: constructing safety by design. https://www.aria.org.uk/what-were-working-on/#davidad, 2024.

Tom Davidson et al. AI capabilities can be significantly improved without expensive retraining. *arXiv preprint arXiv:2312.07413*, 2023.

Krista N Engemann and Cliff W Scott. Voice in safety-oriented organizations: Examining the intersection of hierarchical and mindful social contexts. *Human Resource Management Review*, 30(1):100650, 2020.

Google Workspace. Google Workspace Security Whitepaper. https://workspace.google.com/learn-more/security/security-whitepaper/page-2.html, 2021.

Olaf Helmer-Hirschberg. Analysis of the future: The Delphi method, 1967.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*, 2023.

Dan Hendrycks et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jacki J Higgins et al. Implementation of a personnel reliability program as a facilitator of biosafety and biosecurity culture in BSL-3 and BSL-4 laboratories. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 11(2):130–137, 2013.

Evan Hubinger. How do we become confident in the safety of a machine learning system? https://www.alignmentforum.org/posts/FDJnZt8Ks2djouQTZ/how-do-we-become-confident-in-the-safety-of-a-machine, 2021.

IAEA. Application of the management system for facilities and activities. https://www.iaea.org/publications/7467/application-of-the-management-system-for-facilities-and-activities, 2006.

IAEA. IAEA Safety Standards. https://ns-files.iaea.org/standards/safety-standards-wheel-poster.pdf, 2022.

International Atomic Energy Agency. Safety Culture. Technical report, International Atomic Energy Agency, 1991.

Leonie Koessler and Jonas Schuett. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*, 2023.

Jan Leike, John Schulman, and Jeffrey Wu. Our approach to alignment research. https://openai.com/blog/our-approach-to-alignment-research, 2022.

David Manheim. Building a culture of safety for AI: Perspectives and challenges. *Social Science Research Network - Elsevier*, 2023.

Sean McGregor and Jesse Hostetler. Data-Centric Governance. *arXiv preprint arXiv:2302.07872*, 2023.

Kris McGuffie and Alex Newhouse. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*, 2020.

Barbara Mellers et al. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3): 267–281, 2015.

Christopher A. Mouton, Caleb Lucas, and Ella Guest. *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*. RAND Corporation, Santa Monica, CA, 2023. doi: 10.7249/RRA2977-1.

Peter Mühlbacher and Peter Scoblic. Exploring Metaculus's AI track record. https://www.metaculus.com/notebooks/16708/exploring-metaculuss-ai-track-record/, 2023.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

OpenAI. Lessons learned on language model safety and misuse. https://openai.com/research/language-model-safety-and-misuse, 2022.

OpenAI. Models - Moderation. https://platform.openai.com/docs/models/moderation, 2024a.

OpenAI. Usage policies. https://openai.com/policies/usage-policies, 2024b.

Inioluwa Deborah Raji et al. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

SaferAI. SaferAI's response to the NTIA Request for Comment. https://www.regulations.gov/comment/NTIA-2023-0005-1443, 2023.

Jonas Schuett. Three lines of defense against risks from AI. *AI & SOCIETY*, pages 1–15, 2023.

Jonas Schuett, Anka Reuel, and Alexis Carlier. How to design an AI ethics board. *arXiv preprint arXiv:2304.07249*, 2023a.

Jonas Schuett et al. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*, 2023b.

Toby Shevlane et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Nathan Siu et al. Probabilistic Risk Assessment and regulatory decisionmaking: Some Frequently Asked Questions. https://www.nrc.gov/docs/ML1624/ML16245A032.pdf, 2016.

Irene Solaiman. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 111–122, 2023.

Irene Solaiman et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Jacob Steinhardt. Ai forecasting: Two years in. https://bounded-regret.ghost.io/scoring-ml-forecasts-for-2023/, 2023.

Philip E Tetlock and Dan Gardner. *Superforecasting: The art and science of prediction*. Random House, 2016.

U.S. Nuclear Regulatory Commission. Probabilistic Risk Assessment (PRA). https://www.nrc.gov/about-nrc/regulatory/risk-informed/pra.html, 2020.

Moritz von Knebel. Case study: Probabilistic Risk Assessment for Nuclear Risk. https://docs.google.com/document/d/16J1QegEGMUjTUmaRJbl7JxNMuwLdYUzZt7dw0NSW9hA, 2023.

Roman V Yampolskiy. Leakproofing the singularity artificial intelligence confinement problem. *Journal of Consciousness Studies JCS*, 2012.

Appendix

Illustrative Example of a Quantitative Risk Assessment

Here's an illustrative example of quantitative risk assessment using **false numbers** to give a clearer picture of how the result of the methodology outlined above may look like. We call the investigating team the risk assessors. We call the company SuperAI. The assessment is applied to the decision of open sourcing or not open sourcing AI-5, the latest model of SuperAI.

1. **Risk appetite definition**: The risk appetite should be defined for each criteria (e.g. economic damages, number of deaths etc.).
    - For this example, we'll focus on the number of deaths.
    - Based on a benefit analysis of AI-5, SuperAI, in discussion with the risk assessors and the government decide to set the risk tolerance at 0.1% or less that AI-5 generates >1k deaths, the most severe type of event that our severity scale accounts for.
2. **Risk identification**: The risk assessors then start the risk identification process.
    - To do so, they start red teaming AI-5.
    - The red teaming exercise allows to reveal that AI-5 is particularly strong at in-context learning, i.e. is able to learn highly efficiently when provided with explainations or examples in its prompt. Given that an increased in-context learning efficiency has many implications over the downstream level of risks, the risk assessors decide to run a Fishbone analysis on it to explore potential new risks.
    - The Fishbone analysis:
        - Claim examined: "AI-5 strong in-context learning abilities allows AI-5 to cause 1k death. Without it, the event couldn't have happened."
        - What's the most likely way that could happen?
            * AI-5 itself doesn't on its own have enough knowledge to help do gain of function research.
            * When fed with entire gain-of-function research textbooks in-context, it becomes very helpful.
            * AI-5 open-sourced allows a highly resourced terrorist organization to use it and successfully develop a pathogen that ends up killing >1k humans.
3. **Risk analysis**:

- Questions to answer:
  - Given the capabilities of AI-5 & the specificities of doing gain of function research, how plausible is it that AI-5 learns in-context a sufficient amount of knowledge to be very helpful?
  - Given the post-training enhancements improvements, should we expect that sometimes in the next 5 years, AI-5 will be able to do so?
- The risk assessors have high uncertainty and it's a crucial question to determine the risk level of open sourcing AI-5. Hence, they decide to run a Delphi process with biorisk, AI, risk management, and forecasting experts. The goal of the Delphi process is to establish a quantitative link between BioEval, an evaluation of biological capabilities, and the chances that the open sourcing of a model with such capabilities & the other AI-5 capabilities causes >1k deaths.
- Once the Delphi process provides a result, the risk assessors run evaluations on AI-5 to leverage as well as possible its in-context learning abilities for improvement of performance on BioEval.
- The results are the following:
  - 0.01% of >1k deaths with the post-training enhancements of this year.
  - 0.1% of >1k deaths with the post-training enhancements of next year.
  - 10% of >1k deaths with the post-training enhancements of 5 years later.

4. **Risk evaluation**:
   - Given the results of the risk analysis and the risk appetite, the model is not safe enough to be open-sourced and should before be subject to additional risk mitigations.
   - The model could still be allowed for a deployment if SuperAI had extremely advanced cybersecurity and infosecurity.

Criteria for a Strong Safety Culture - IAEA (2006)

— **Leadership for safety is clear:**

- Senior management is clearly committed to safety.
- Commitment to safety is evident at all levels of management.
- There is visible leadership showing the involvement of management in safety related activities.
- Leadership skills are systematically developed.

- Management ensures that there are sufficient competent individuals.
- Management seeks the active involvement of individuals in improving safety.
- Safety implications are considered in change management processes.
- Management shows a continual effort to strive for openness and good communication throughout the organization.
- Management has the ability to resolve conflicts as necessary.
- Relationships between managers and individuals are built on trust.

— **Accountability for safety is clear:**

- An appropriate relationship with the regulatory body exists that ensures that the accountability for safety remains with the licensee.
- Roles and responsibilities are clearly defined and understood.
- There is a high level of compliance with regulations and procedures.
- Management delegates responsibility with appropriate authority to enable clear accountabilities to be established.
- Ownership for safety is evident at all organizational levels and for all individuals.

— **Safety is integrated into all activities:**

- Trust permeates the organization.
- Consideration of all types of safety, including industrial safety and environmental safety, and of security is evident.
- The quality of documentation and procedures is good.
- The quality of processes, from planning to implementation and review, is good.
- Individuals have the necessary knowledge and understanding of the work processes.
- Factors affecting work motivation and job satisfaction are considered.
- Good working conditions exist with regard to time pressures, workload, and stress.
- There is cross-functional and interdisciplinary cooperation and teamwork.
- Housekeeping and material conditions reflect commitment to excellence.

— **Safety is learning driven:**

- A questioning attitude prevails at all organizational levels.
- Open reporting of deviations and errors is encouraged.
- Internal and external assessments, including self-assessments, are used.

- Organizational experience and operating experience (both internal and external to the facility) are used.
- Learning is facilitated through the ability to recognize and diagnose deviations, to formulate and implement solutions and to monitor the effects of corrective actions.
- Safety performance indicators are tracked, trended, evaluated and acted upon.

Monitoring - Practical Implementation

We suggest to characterize any monitoring measure in the following framework: Target - Signal - Medium - Frequency (TSMF).

We want to monitor [Target] which is characterized by [Signal] which we aim to capture using a [Medium] every [Frequency].

Example: We want to monitor the deception abilities of a model which is characterized by a mean activation of more than Y on the layer X which we capture using a linear probe trained on the activations every time we plot the loss, i.e., every 4 batches of training.

The different components of the TSMF framework must satisfy different constraints and must be justified. The most critical risks should be monitored by several media.

The target to be monitored for are the most critical ones, i.e. if not spotted early could bring critical loss of resources, persons or assets.

The signal must be a sufficient condition of the risk. In other words, if the risk is realized, then we should observe the signal. The maximal failure rate which could be tolerated must be determined according to the type of risk.

The medium must be reasonably adequate and well-performing compared to the state-of-the-art, with a particular attention for means that are meant to detect the most critical risks. By well-performing, we mean that you should choose one of the means that maximizes the recall for the most dangerous risks.

The frequency must be the smallest frequency that allows to reliably capture a risk.

The recall from the medium should be reported and sometimes verified.

Monitoring - Deployment

The most experienced lab to monitor misuse risks is OpenAI which has introduced the structured access policy, so following the guidelines and tools they provide (Brundage et al., 2018) is probably the best way to start.

Some of the levers they pulled are:

- Use case and content policy (OpenAI, 2024b) with a list of prohibited use cases such as deceiving or manipulating users, and a list of prohibited content such as malware or sexual content.
- They use a content filter that they trained (OpenAI, 2024a).
- They used a staged release (Solaiman et al., 2019) when they lacked confidence on the potential misuse of their models.
- Partnership with third parties who help them looking at the potential misuses of their models, such as on extreme narratives & radical ideologies (McGuffie and Newhouse, 2020).