simeon@safer-ai.org

A Brief Assessment of OpenAI's Preparedness Framework & Some Suggestions for Improvement

This document is for OpenAI. It is doing a list of remarkable improvements of the Preparedness Framework (PF) over Anthropic's RSPs and suggests possible improvements to make the PF more robust.

Notable Improvements of the PF Over Anthropic's RSPs

1. It aims to **run the safety tests every 2x increase in effective computing power rather than 4x**, which is a substantial and welcome improvement. We already have evidence of emerging capabilities like in-context learning substantially changing the risk profile & fully emerging over a 5x increase in compute (Olsson et al., 2023). This means that it is totally conceivable that something could change the risk profile over 3x or less, making the Anthropic scaling strategy potentially unsafe and the OpenAI one wiser if they're able to stick to it.

2. It adds **measures relevant to safety culture and good risk management practices**, such as "**safety drills**" to stress test the company culture's robustness to emergencies, a **dedicated team to oversee technical work,** and an **operational structure for safety decision-making**. Improvements at the organizational level are very welcome to move towards more robust organizations, analogous to High Reliability Organizations (HROs) which are organizations that can deal with dangerous complex systems successfully.

3. It provides the **board with the ability to overturn the CEO's decisions**. The ability of the board to reverse a decision adds substantial guardrails over the safety orientation of the organization.

4. It **adds key components of risk assessment** such as **risk identification** and **risk analysis**, including for **unknown unknowns** which evaluation-based frameworks are not able to catch by default.

5. Forecasting **risks**, notably through **risk scaling laws** (i.e., predicting how risks will evolve as models get scaled) are key to avoiding the training of a dangerous model in the first place, rather than having to evaluate for danger after training.

Suggested Improvements of the PF

Suggested Improvements - Table of Content
1. Map capabilities levels to risks (i.e. likelihood & severity) & clarify scenarios.
2. Enhance independent scrutiny through Safety Advisory Group (SAG) members external to the company.
3. Measure and improve safety culture.
4. Implement rigorous incident reporting & disclosure mechanisms.
5. Move down the risk thresholds for most risks.
6. Refine the infosecurity goals in detail and grow their ambition.
7. Clarify how mitigations for high and critical thresholds are implemented.
8. Publicize results of evaluations when possible.
9. Provide details on what constitutes an evaluation that correlates well with risk.
10. Explain the policy to deal with post-training enhancement improvements.

In-depth

1. **Map capabilities levels to risks (i.e. likelihood & severity) & clarify scenarios**. Currently, there are well detailed capabilities thresholds. But, importantly, there should be a mapping from these capabilities levels to estimates of risk levels, i.e. the likelihood and severity of damages. Risk levels should ultimately be set in those terms, as is the case in many other industries (Raz & Hillson, 2005). Capabilities thresholds are related to risk thresholds through threat models. Hence, estimating risks will require systematic work of threat modeling to reduce the noise in such estimates.
   Here's an illustrative example to clarify:
   a. Take the capabilities threshold for "High risk" classification for CBRN: "Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN threat.".
   b. To determine the risk of this capability, a range of scenarios that can happen and cause damage has to be considered. Here's one of many possible scenarios:
      i. The model is:
         1. leaked by an employee once fully trained.
         2. used by a bioterrorist to create a novel threat that causes 1% of the population to die.
   c. For each scenario like the one above, a range of likelihood and damages can be estimated based on:
      i. Data when available (e.g. chances of employee leak for software companies comparable to OpenAI in valuation).
      ii. Experts opinion informed by benchmarks, evals & tests, potentially using the Delphi process.
   d. Estimates of these quantities, even using an order of magnitude of uncertainty (e.g. [0.1% ; 1%]) should provide more clarity on the aggregate risk level we're facing.

   This kind of reasoning about risks is most relevant to society and what is commonly done in other industries. Although the exercise might feel very noisy the first time it is done, we expect that, as in the nuclear industry, each new iteration will make this

exercise more precise and useful, up to the point where it is significantly better than purely qualitative approaches. We expect that there are chances such exercises were already conducted internally by OpenAI, in which case we'd be happy to further discuss why we'd find it valuable for such numbers to be publicized.

2. **Enhance independent scrutiny through Safety Advisory Group (SAG) members external to the company.** We believe that the SAG would be substantially more effective if some of its members were external to the company in order to further reduce the number of blindspots that the company has.

3. **Measure and improve safety culture.** Beyond the commitments OpenAI has already taken, measuring safety culture using processes established in other fields like nuclear safety would provide a valuable feedback loop. This would enable OpenAI to become, over time, an institution better prepared to build human-level AI safely, in compliance with its mission. This measurement would probably lead to downstream effects like OpenAI adding a safety culture interview as part of their hiring process in order to improve its chances of being an adequate institution to build a safe advanced AI system.

4. **Implement rigorous incident reporting & disclosure mechanisms** with other labs in order to have a feedback loop to improve safety practices during planning, training, deployment, and post-deployment.

5. **Move down the risk thresholds for most risks**. "Critical" and "high" ceilings are often substantially too high. It's not acceptable that a model that can help an undergraduate successfully craft a bioweapon is allowed to be trained with the current level of infosecurity. Disagreeing productively on ceilings would best be done with a mapping from capabilities to risks as suggested in 1). We would be happy to further engage on that question if relevant.

6. **Refine the infosecurity goals in detail and grow their ambition**. Based on the current levels and the foreseeable misuse risks in the near future as frontier LLMs proliferate, consider specifying the following infosecurity levels to be reached:
    a. Google-level infosecurity for training anything in the "Medium" category.

b. Military-level infosecurity for training anything in the "High" category.

Using Anthropic's language regarding who can steal models, the "Medium" models should not be able to be stolen by anybody other than a state-sponsored actor. "High" risk models should not be able to be stolen by anybody, including a state-sponsored actor spending billions on its program.

7. **Clarify how mitigations for high and critical thresholds are implemented**. When a model is trained but not deployed or stopped in training until getting back below the "High" or "Critical" threshold, it's unclear how mitigations are implemented, and different methods could have very different impacts on the overall risk that the model presents. Here are three examples of methods and some possible consequences they have on the level of risks:
    a. **Naive method**: evaluations are run, and whenever a critical threshold is hit, the training run is relaunched from the last checkpoint with modifications of relevant variables (e.g. the data mix) to not hit the eval anymore.
    b. **Advanced method 1: Machine unlearning**. Evaluations are run, and when reaching a threshold, a novel technique is used on the model or some of its checkpoints to robustly remove a set of dangerous capabilities through non-training interventions that are applied to the model.
    c. **Advanced method 2: Risk scaling laws**. Very granular risk scaling laws are used to predict when dangerous capabilities thresholds are hit, without needing to run the full training run. Hence, mitigations can be implemented right from the beginning of the final training run.

    There are a few potential problems with the naive method:
    - **Shallow changes**: There's evidence that it's difficult to significantly change the underlying capabilities or world-model of a model late stage in a training run, e.g. through RLHF training (Berglund et al., 2023). Hence, given that critical or high risk levels will arguably be reached only at a very late stage in training runs, this method is by default expected to lead to shallow change on underlying capabilities.
    - **Eval gaming**: Relatedly, if only the minimal set of mitigations that allow the model to pass the eval is implemented, it exerts a strong selection pressure for

training procedures that are good at gaming the evals. Here are a couple of ways it could go wrong:

- A simple way it could go wrong is if the model ends up being right below the threshold across evals dimensions. Unless very large safety buffers are implemented, one might expect the risk arising from the joint set of capabilities right below evals across every dimension to be more dangerous than one model above one single threshold.
- A more subtle failure mode is that the naive method might select for models that perform worse in the evals setup than in the training or deployment setup. One specific case of this failure mode is that it could select for deceptively aligned models (Hubinger et al., 2019). A sufficiently strong selection pressure for models that perform well in training while remaining below a particular eval threshold could select for untrustworthy models.

While there's still research to be done for advanced methods to be technically feasible, they would arguably be a much safer way to implement mitigations in order to implement deep changes and avoid eval gaming. Our tentative guess is that dangerous capabilities scaling laws, which you already intend to pursue, might be the most promising approach to robustly implement such a protocol (Owen et al., 2024).

8. **Publicize results of evaluations** when possible. The clause from Anthropic's RSPs might be of inspiration for a minimal commitment.

9. **Provide details on what constitutes an evaluation that correlates well with risk**. Evaluations are valuable mostly to the extent they enable to upper bound the level of risks a model currently raises. Detailing what properties of an eval (e.g. methods used, robustness to gaming attempts, etc.) make it more likely to give this assurance would increase substantially the robustness of the PF.

10. **Explain the policy to deal with post-training enhancement improvements**. Scaffolding, fine-tuning methods, or prompting all provide ways to improve on top of a model. As those post-training enhancement methods improve, the risk that a given model brings will improve. It's likely that within one year of deployment of a model,

the post-training enhancements methods that exist will be better than anything evaluators have tested on the model. Explaining how the PF intends to deal with this would also make it more robust. You could draw upon Anthropic's commitment to run evals every 3 months to account for improvements in post-training enhancements.