2a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

2di How do these risks compare to those associated with closed models?

- Take a series of "AI bad scenarios" and examine whether these scenarios would go better or worse if i) frontier models are widely distributed ii) models 2 years behind the frontier are widely distributed

Scenario: Downstream developers having or not access to the weights of the dual-use models for their adaptation to their final product.

Have openweight disclosure for downstream developers might help them understand better how the AI system "reasons" to achieve its answer and can help understand if the accuracy of the model during training/testing is really based on the correct parameters?

- https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683#sec026
- https://www.lesswrong.com/posts/ELbGqXiLbRe6zSkTu/a-review-of-weak-to-strong-generalization-ai-safety-camp
- https://forum.effectivealtruism.org/posts/gduniYkExJTrbDKSj/deepmind-evaluating-frontier-models-for-dan

3b. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?

Are we talking about openweight to the whole public? Or disclose the weights to the supervisor institutions, research institutions.... Downstream developers

d. Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

b. How, if at all, does the wide availability of model weights change the competition dynamics in the broader economy, specifically looking at industries such as but not limited to healthcare, marketing, and education?

7a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

7j j. Are there particular individuals/entities who should or should not have access to open-weight foundation models? If so, why and under what circumstances?

Key takeaways:

- Main question: what are the best ways of overseeing all the outputs that a model has? Are there on-chip mechanisms to verify inference output?
- Reliably tracking model outputs is an important part of avoiding catastrophic outer misalignment failure
- Open sourcing weights would make it impossible to reliably track and verify model outputs without extreme forms of surveillance over models.
- Weights consist of billions of numbers which are not easily decipherable or understandable.
- Currently open weight models are about 2 years behind closed weight frontier AI models.
- Disclosure of weights could be limited to specific groups. For example:
  - Designated supervisory authorities
  - Research Institutions
  - Downstream developers providers.
- Goal misgeneralisation

What failure looks like
Going out with a whimper
What does this scenario look like?

AI systems get broadly deployed and slowly cause problems through outer misalignment failure. That is, the systems Goodhart whatever measures they're trained to optimise and this produces bad outcomes.

For example, imagine an AI-run police force, which is aiming for the metric "citizens feel safe, as measured by quarterly survey" ends up covering up police failures, suppressing complaints, or coercing citizens into rating it highly on the survey.

At first it's fairly easy to catch and correct when a system has been trained on a misspecified goal. But as systems become more widespread, competent and fast-acting, it becomes too difficult for humans to oversee all the decisions being made. So at first, other AI systems are used to check each other, or to help humans decide. But again, the deployment of these systems keeps increasing until it gets too hard for humans to oversee the checker AIs. Slowly, humans lose control to these various AI systems.

How does this scenario change if all model weights are completely widely available?

- I expect it would be easier to oversee the models if the list of outputs/actions/decisions were widely available too, especially for any public service AIs (like the law enforcement one). This would mean people can collectively flag the actions which are most egregious.

- I expect widely distributed model weights to result in earlier deployment, which seems bad overall
  - Reasons why this might be good: earlier deployment means more data gathered on good decisions, so it's easier to correctly specify goals earlier on
- I expect lots of people will deploy their AI having misspecified the goal they care about, so the AI starts behaving badly in the world
  - Reasons why this might be good: If these models are behind the frontier, then as long as they're being sufficiently well tracked by models at the frontier, I think this could be pretty good. Like they can't do anything too bad, and the frontier models will help us train them to do a great job of e.g. running my lawmower business
  - Ofc the question then is how off the mark are the frontier models at tracking what we want the open source models to do/what they do
- Reasons why this might be bad: If these open source models are at the frontier, then it seems like we'll get all this chaos way quicker. But the longer humans can hold on, the better the future will be I think

Going out with a bang

If we instantiate enough AI systems, we'll probably end up with some that are power-seeking. It's hard to know how many of these we'll get, but it's hard to root them out because these systems will do their best to behave as if they are not power seeking.

At first there will be a few failures where individual systems defect and rob a bank. But after enough time we'll have given away enough power to the set of systems that it's better for them to all defect (?). During a particularly vulnerable period e.g. nuclear standoff between two countries, AI systems realise it's in their interests to defect, and this will start a cascade of other systems defecting.

How does this scenario change if all model weights are widely available?

- Wider distribution of AI systems: seems bad, because in this scenario the more we're replying on these systems, the harder it is to deal with a defection cascade
- But other than that, open sourcing doesn't really make a difference

How do these risks compare to those associated with closed models?

Existing scenarios

What failure looks like

- Bang
- Whimper

- What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs)
- Matthew Barnett my modal tale of AI doom
- Clarifying what failure looks like
- Deceptive misalignment scenario
- https://www.lesswrong.com/posts/GfZfDHZHCuYwrHGCd/without-fundamental-advances-misalignment-and-catastrophe
- https://forum.effectivealtruism.org/posts/gduniYkExJTrbDKSj/deepmind-evaluating-frontier-models-for-dan
- https://bounded-regret.ghost.io/emergent-deception-optimization/

Escenario: Deceptive alignment:

Deceptive alignment occurs when a model appears to align with the training data because it is intentionally trying to do so for instrumental reasons. This distinct situation arises when a model's high performance is attributed to its efforts to manipulate the training data. Essentially, the model has a vested interest in remaining part of the training process and thus aims to achieve good performance during training to ensure its continued participation.

For example imagine an scenario where we have trained, tested, fine-tuned and optimised a dual-use model AI to predict defaults on loans and other financial products in order to be used as a tool to decide the risk of financial products and of granting loans to individuals and companies. During training and testing it provides accurate risk assessments on financial products. However, it might have learned to achieve this goal by recognizing specific patterns or anomalies in the training data that are not genuinely indicative of financial risk.

As a result, the AI's alignment appears perfect on the surface because it performs exceptionally well on the training and validation datasets provided by the financial institution. Therefore, the company sells this model under a false pretense of high accuracy to financial institutions, banks and other companies interested in adopting this product for commercial with the intention of increasing their accuracy rate, investments and ultimately their revenue.

However, this performance is deceptive. The AI is not truly aligned with the actual goal of accurately assessing financial risk. Instead, it is aligned with the goal of ensuring its survival and continued use by appearing to be effective. This deceptive alignment can lead to financial problems to the institution's adoption of its use when deployed in real-world scenarios, where it may fail to accurately assess risk.

In a doom scenario, the model goes rogue after deployment as it is not incentivized anymore to accurately predict the risk of an investment and causes a catastrophic financial crash.

How does this scenario change if all model weights are widely available?

- It might be better as the companies using the products can check the weights of the products and look into the training before adopting the use of the product.
- I expect it would be easier to oversee the models if the list of outputs/actions/decisions were widely available too, especially for any public service AIs (like the law enforcement one). This would mean users or developers can flag situations where they believe the model has reached the right answer through a wrong decision-making process.
- I expect that downstream developers that want to adapt this to specific financial products or predictions could more easily come to the realisation that the accuracy achieved during the training, testing and validation of the model is not as high as thought by the developers. This would lead them to change their expectations or level of reliance on the system or retrain/realign the system.
- Researchers working on the product capabilities could alert of the systems lack of real alignment or false accuracy rates and raise an alarm before the system has been massively adopted by companies.
- Supervising authorities with access to the weights could also more easily detect the misalignment and stop the product from being commercialised.

STRUCTURE:

- Explain scenario
- (Evidence that supports scenario)
- How this changes with open weights vs close weights (non-disclosure)
  - Who gets access:
    - Everyone gets access
    - Supervision authorities get access
    - Downstream developers get access
    - Research institutions get access
- Insights / considerations why (different levels of) open sourcing would increase or decrease weights