This doc collects some notes on safety evals in the context of slowing dangerous AI research. Some of my thoughts here were inspired by a document by Jide Alaga.

My overall view:

- I consider safety evals a promising intervention for slowing the development of dangerous AI. I think there's a significant risk that poorly-implemented evals will do harm – but that just means we ought to be careful in how they're set up and who has the authority to approve auditors.

A framing that I like:

- We can frame safety evals as an attempt to make warning shots happen in a controlled environment.
- Auditors publishing the results of their audits (and salient examples like AI hiring a task rabbit to get through a captcha) are pretty helpful for creating awareness of specific risks. (There's the issue of audit logs getting into the training data of future AI models, making them better at gaming audits.)

How to get labs to agree to be audited:

- I think the AI safety community has a lot of leverage here. If people who have a track record of caring about AI safety before it became mainstream get together and condemn the big labs for not even agreeing to a common-sense auditing scheme, the labs would look really bad. (If anything, I'm maybe more concerned that the AI safety community might give up too much leverage if they signal that things are fine as long as the labs agree to audits – but this depends on specifics of how thorough the audits would be and the consequences if an audit fails, etc.)
- **I think it's important for evals to be tied directly to a moratorium of sorts that labs have committed to in advance (or that's backed by legislation).** I think it would be unfortunate if evals are on a voluntary basis – I think that would be the safety community giving up too much leverage.

On the importance of good auditing standards:

- Comment on the idea of multiple auditors:
  I'm concerned that as soon as there's an appearance of "safety researchers disagree about their threat models" (e.g., "some think cooperativeness is important, others focus on deceptive alignment"), this makes it easier to altogether dismiss safety-related concerns with moves like "these concerns seem speculative" or "we've addressed one of the concerns, so we should be fine – we don't agree with the other one." I think it's important for the safety community to (1) reach a consensus about minimum safety standards they think AI systems have to pass, and (2) the consensus should reflect *what's actually risky* about advanced AIs, rather than provide a sense of false security.
- Having multiple auditors or disagreements among auditors increases the chance that some auditors are incompetent or corrupt. Bad auditors are arguably worse than useless (my sense is that large-scale financial scams were almost never uncovered by auditors – instead, they had incompetent or corrupt auditors giving them cover). I worry that there are social incentive gradients for auditors to be very accommodating toward the powerful AI labs. More auditing options increases the risk that some auditors will go too far down these gradients.

Evan Hubinger in Towards understanding-based safety evaluations:

- "Understanding as a safety standard also has the property that it is something that broader society tends to view as extremely reasonable, which I think makes it a much more achievable ask as a safety standard than many other plausible alternatives. I think ML people are often Stockholm-syndrome'd into accepting that deploying powerful systems without understanding them is normal and reasonable, but that is very far from the norm in any other industry. Ezra Klein in the NYT and John Oliver on his show have recently emphasized this basic point that if we are deploying powerful AI systems, we should be able to understand them."
- If Evan's view here is correct, there's a risk that the wrong kinds of safety evals (ones that don't rely on understanding) would mostly give us a false sense of security.
    - They'd still help catch model behavior that's outright misaligned and dangerous, so there is a significant potential upside. Whether that upside is worth it depends on our priors on deceptive alignment and how hard it would be to catch. (Evan's stance being: we may not be able to catch it with safety evals.)

On benchmarks potentially being net negative:

- Some commenters are concerned that safety evals could be negative if they come with helpful benchmarks that ML practitioners can optimize against, boosting capabilities. This particular objection could maybe be addressed by not having legible benchmarks – instead, auditors simply give their all-things-considered judgment about whether the model passes. (Comparison to tests at school: don't give out last year's tests, and formulate new understanding-driven questions for each new test set, as opposed to using questions from a database.)