IAPS | Institute for AI Policy and Strategy

11th March - 2024

Responsible Scaling: Comparing Government Guidance and Company Policy

AUTHORS

Bill Anderson-Samways - Research Analyst

Shaun Ee - Policy and Strategy Manager

Joe O'Brien - Research Analyst

Marie Buhl - Research Analyst

Zoe Williams - Acting Co-Director

Executive Summary

As advanced AI systems scale up in capability, companies will need to implement practices to identify, monitor, and mitigate potential risks. "Responsible capability scaling" is the specification of progressively higher levels of risk, roughly corresponding to model size or capabilities, and entailing progressively more stringent response measures. We evaluate the original example of a Responsible Scaling Policy (RSP) – that of Anthropic – against guidance on responsible capability scaling from the UK Department for Science, Innovation and Technology (DSIT).

In October 2023, DSIT produced a list of best practices for responsible capability scaling, the most detailed guidance available from an authoritative source at present. In the lead-up to the UK AI summit in November 2023, independent analysis from the Leverhulme Centre for the Future of Intelligence found that, out of all AI companies, Anthropic performed best on responsible capability scaling. Building on the Leverhulme Centre's piece, we analyze Anthropic's RSP measures more deeply against a selection of best practices from DSIT guidance, using our interpretation of DSIT's guidance as well as relevant literature.

Although we think that Anthropic's RSP is a commendable first step to manage advanced AI risks, we identify areas of improvement when it comes to fulfilling several important DSIT practices. On that basis, we make several recommendations for improving RSPs, not just for Anthropic but for other AI companies and government standard-setting bodies. Our top recommendations are:

- **Anthropic and other AI companies should define verifiable risk thresholds for their AI safety levels (ASLs - or equivalent), informed by tolerances for "societal risk" (SR) in other industries. Such risk thresholds should likely be lower than Anthropic's current thresholds, and should be defined in terms of absolute risk above a given baseline, rather than relative risk over said baseline.**
    - SR tolerances are risk-tolerances for events involving multiple fatalities.
    - A non-exhaustive survey by Flamberg et al. (2016) suggests that "maximum" SR tolerances for events involving ≥1,000 fatalities – Anthropic's definition of a "catastrophic risk" – range between 1 E-04 to 1 E-10 such events per year.[1] "Broadly acceptable" tolerances are generally two orders of magnitude lower.
    - We tentatively suggest that companies set their ASL-4 threshold (or equivalent) in the "maximum" SR range, and their ASL-3 threshold (or equivalent) two orders of magnitude lower than their chosen ASL-4 threshold.[2] We think that

[1] That is equivalent to a range of 0.01% to 0.00000001% per year in terms of *probability.*

[2] Importantly, the SR threshold applies to pre-mitigation, not post-mitigation, risk. Thus, if a company selected the 1 E-04 figure for their ASL-4 threshold (or equivalent), a model would be classed as ASL-4 if its pre-mitigation risk exceeded that 1 E-04 threshold.

Anthropic's current risk thresholds probably exceed the maximum SR range.
    - Anthropic's current risk thresholds also sometimes employ a relative rather than absolute definition of risk – for example, requiring AI misuse risk to be roughly equivalent to the non-AI misuse risk baseline (Anthropic's tentative ASL-4 misuse threshold). However, the baseline risk could fluctuate, and is highly uncertain. Using an absolute risk threshold, such as the SR tolerances defined above, would be better.[3]
    - Qualitative risk thresholds could also be used in tandem with the above.
    - **Ultimately, however, a government body, such as UK DSIT or the US National Institute for Standards and Technology (NIST), or an industry body such as the Frontier Model Forum (FMF), should develop standardized operationalizations of risk-thresholds for RSPs.**
- **Anthropic and other companies should also specify SR thresholds for a granular set of risk *types* at a given safety level** – for example, not just "misuse" but "biological misuse" as opposed to "cyber misuse."[4]
- **Anthropic and other companies should detail when they will alert government authorities of identified risks**. Currently, Anthropic's RSP does not mention communication with governments outside of a narrow case (involving Anthropic's response to a bad actor scaling dangerously fast). We suggest that risks should at minimum be communicated to relevant agencies when they reach a given threshold, for example the ASL-3 or ASL-4 thresholds outlined above. Upcoming work from IAPS on coordinated disclosures will explore this question in greater depth.
- **Anthropic and other companies should commit to external scrutiny of both their evaluation methods** (i.e., whether those methods work) **and their individual evaluation results at ASL-3 or sooner.**

In addition, we note that Anthropic's RSP performs well on many of the UK government practices that we think are very helpful for reducing catastrophic risks. It is important that other companies take similar measures, especially in areas where inter-company coordination is needed, such as planning for the need to pause model development if adequate risk-mitigation measures are not in place.

---

[3] Of course, that is not to say that risk thresholds should not be defined with respect to a baseline. They should simply be defined in terms of absolute risk above a given baseline - e.g., "a 0.01 percentage-point chance of a biological catastrophe caused by AI (i.e., a catastrophe that would not have been possible on the non-AI baseline)."

[4] Anthropic has committed to a work plan to develop domain-specific risk evaluations for autonomous replication and CBRN and cyber threats, however does not currently include detail on how ASL boundaries for these will be specified in regard to societal risk / outcomes. For instance, their model card for Claude 3 includes separate evaluations of three capabilities: Autonomous Replication and Adaptations, Biological, and Cyber. Each domain has specific ASL-3 boundaries. The autonomous replication and adaptations boundary is 'the model passing 50% or more of the tasks described below with at least 10% pass rate (i.e. succeeding at one in ten attempts)', which is difficult to quantify in terms of societal risk.

Below, we provide our full recommendations, this time grouped sequentially in terms of the practices recommended by DSIT. DSIT's list includes 27 best practices; through an internal ranking exercise, we identified seven as especially important for an RSP that mitigates the largest risks from advanced AI.

DSIT-recommended practice	Our recommendations and reasoning
Developing rigorous **risk assessment processes** for models	Following Anthropic's example, **companies should conduct both evaluations of current capabilities and forecasts of future capabilities.** (read more) **When developing evaluations of advanced AI models, companies and governments should consider how to account for ways in which evaluations may systematically underestimate the capabilities of models**, for example, the possibility that models may face incentives for "deceptive alignment," appearing aligned when in fact they are not, or the fact that evaluations only provide an example of dangerous capabilities, not the upper bound. To ensure accurate risk assessments for advanced AI systems, **companies and governments should invest as needed in targeted measures such as mechanistic interpretability research, evaluations of deceptive alignment, and other measures that could reduce the likelihood of under-estimating model capability.**
Monitoring systems **throughout development and deployment**	**Companies should monitor AI systems both pre-training, during training, and post-deployment (rather than only pre-deployment).** **For example, Anthropic's RSP currently contains decently strong monitoring measures during *training*, but its monitoring measures *post-deployment* are limited to automated vulnerability detection.** At higher ASLs, Anthropic could also commit to (for example) implementing incident monitoring measures from cybersecurity such as 24/7 security operations centers. **On a related note, throughout the lifecycle, AI developers should combine automated vulnerability detection (for**

	example, input-output monitoring) with human-led vulnerability detection. (read more)
Describing, operationalizing, and continually refining **risk thresholds**	**Anthropic and other AI companies should commit to somewhat lower, more specific risk-thresholds, defined in terms of absolute risk (e.g., expected frequency of catastrophic events above a baseline) rather than relative risk (proportional increases over a baseline).** (read more) • Governments and industry associations often set lower risk tolerances for "societal risk" (SR) – i.e., single events involving many fatalities – than for isolated individual fatalities, capturing the intuition that the latter will often be more difficult to prevent. • A non-exhaustive survey across countries and industries by Flamberg et al. (2016) suggests that "maximum" SR tolerances for events of 1,000 or more deaths (Anthropic's definition of a catastrophic risk) range from 1 E-04 to 1 E-10 such events per year. "Broadly acceptable" risk-tolerances are generally set two orders of magnitude lower. • We very tentatively suggest that Anthropic's risk-threshold for ASL-4 (its highest safety level) should be in a similar range to said maximum SR tolerances, and that its threshold for ASL-3 should be around two orders of magnitude lower. The same should apply to other companies' equivalents of ASL-4 and ASL-3. • Those SR tolerances are probably lower than Anthropic's current risk-thresholds. For example, Anthropic's (tentative) ASL-4 threshold for misuse risk is that AI misuse risk in a given domain (e.g., biosecurity) should equal the baseline non-AI misuse risk in that domain. However, research suggests that the baseline risk for a catastrophic biosecurity event may be around 1 E-02 events per year on average, two orders of magnitude above the upper bound for SR tolerance in other industries. • Anthropic's use of a relative risk threshold for ASL-4 ("equivalent to baseline risk") lacks sufficient clarity, because (a) the baseline risk could fluctuate, and (b) different studies will suggest different baseline risks, leaving a risk of companies "picking and choosing" baseline risks. An absolute risk-threshold, such as the SR thresholds

	suggested above, would be better. • Qualitative risk thresholds could be used in tandem with SR thresholds, as some risks may be difficult to quantify even via subjective judgment. • **However, the above are only very early recommendations. We think that a government body, such as UK DSIT or NIST, or an industry body such as the FMF, should commit to a work plan for eventually developing standardized operationalizations of risk thresholds for RSPs.** **Moreover, Anthropic's breakdown of risk *types* within each risk tier is too high-level.** Anthropic has committed to a work plan for better defining threats and evaluations for autonomous replication, cyber and CBRN threats. We suggest that other companies also commit to this, and that Anthropic and other companies ensure their scaling plans link the operationalization of thresholds developed (e.g. via specific capability evaluations) to societal risk tolerances as outlined above. That will help ensure that risk management practices are sufficient for the level of risk and targeted towards the specific risks at hand.
Committing to only proceed with development / deployment **if certain mitigations are in place**	**Following Anthropic's example, companies' risk-mitigation actions should not just be procedural (i.e., follow a predefined set of steps), but should continue until risk is reduced beneath a pre-specified threshold.** (read more)
Informing relevant government authorities once a risk threshold has been crossed	**Anthropic and other companies should adopt a specific threshold above which they commit to alerting relevant government authorities of a given risk.** (read more) • Currently, Anthropic's RSP largely does not discuss government reporting requirements, except in the case of Anthropic consulting with governments on suspending their RSP restrictions to avoid an "imminent global catastrophe" wherein a bad actor is scaling their AI model dangerously quickly. However, that represents a high threshold for harm and a limited set of circumstances. • **An appropriate threshold would be specific and trigger at a lower level than "imminent global**

	catastrophe," for example the tentative ASL-3 or ASL-4 thresholds that we propose above.
Preparing to pause training runs / reduce model access	**Anthropic's commitment to pause development if adequate risk-mitigation measures are not in place should be adopted by other companies.** (read more) • Companies and governments should consider whether there is any additional infrastructure (e.g., design of Service Level Agreements) that needs to be set up to manage potential pauses.
Including **verification mechanisms**	**Anthropic and other companies should examine external risk-verification mechanisms in more detail, and said mechanisms should kick in earlier.** (read more) • For example, committing to getting individual evaluation *results* verified by external auditors rather than just the evaluation methods. • As a first step, Anthropic should commit to external verification of all evaluation results at ASL-3 rather than ASL-4, and other companies should follow their example.

Table of Contents

Introduction

In recent months, "responsible scaling policies" (RSPs) have become a prominent element of some key AI governance frameworks, including that of the UK government. The basic concept behind RSPs is the specification of progressively higher risk tiers, roughly corresponding to model size/capabilities and entailing progressively more stringent response measures. But how do existing companies' RSPs hold up to government guidelines?

A previous review by Ó hÉigeartaigh et al. (2023) used expert judgment to compare major companies' AI safety policies to broad guidance from the UK Department for Science, Innovation and Technology (DSIT). To take a more granular approach, this issue brief zooms in on RSPs, restricting our analysis to the original example of an RSP laid out by Anthropic and comparing it with DSIT guidance on responsible scaling specifically. Whereas Ó hÉigeartaigh et al. could necessarily only examine DSIT's 42 higher-level recommendations (including on RSPs), narrowing our scope allowed us to evaluate performance against DSIT's more concrete sub-recommendations on RSPs.[5]

We identify seven DSIT sub-recommendations that we think would be vital to any RSP's success at reducing catastrophic risk[6] (see below). We then use our judgment to assess how far Anthropic's RSP follows said recommendations (see below). Compared to Ó hÉigeartaigh et al., our analysis is more qualitative; we do not quantify our judgments. The purpose of this issue brief is not to single out Anthropic for criticism – far from it. Instead, we examine Anthropic's RSP precisely because it represents a compelling and brave first step toward addressing advanced AI risks, and Anthropic's concrete RSP commitments allow us to draw more general lessons about how other companies can implement DSIT's suggested best practices.

What is Anthropic's RSP?

Anthropic's RSP is not the only example of a Responsible Scaling Policy. OpenAI has also laid out a "Preparedness Framework" (OpenAI, 2023), which tracks several risk categories and establishes safety baselines. However, Anthropic's RSP is the *original* example of an RSP, so we limit ourselves in this section to outlining Anthropic's RSP.[7]

---

[5] For context, whereas an example of a higher-level DSIT recommendation is "Make preparations to pause development and/or deployment", a more concrete sub-recommendation would be "Prepare to pause training runs or reduce access to deployed models…"

[6] Here, we focus primarily on safety and security risks from advanced AI models. While we consider other risks also important, we did not explicitly consider these in our analysis, though we imagine that an RSP that followed all our recommendations would likely also be effective at addressing many other risks.

[7] Other actors besides companies have also discussed RSPs; for example, UK DSIT's guidance on emerging processes (discussed further below), and METR's description of RSPs (METR, 2023).

Anthropic's RSP framework is modeled after the international standards for biological laboratory safety known as the Biosafety Levels (BSLs). Like the BSLs, Anthropic's RSP is a tiered system, with thresholds at each higher tier triggering new response measures (on top of those already specified at the lower tiers). And like the BSLs, Anthropic has so far specified four such tiers, which it terms "AI safety levels" or ASLs (although Anthropic stresses that it could plausibly specify further ASLs in the future).

Each ASL is triggered if Anthropic's evaluations of model capabilities reveal that a model (either during training or post-training) possesses certain dangerous capabilities. At each level, there are then two buckets of response measures – "containment" measures, which deal with risks that arise from simply *possessing* a given AI model (for example, the risk that a model may be stolen and misused); and "deployment" measures, which deal with risks that arise from the deployment of AI models by the company (for example, harms caused by model users querying an API). The below table outlines Anthropic's commitments thus far – in terms of evaluations thresholds, containment measures and deployment measures:

AI Safety Level	Dangerous Capabilities	Containment Measures *Required to store model weights*	Deployment measures *Required for internal / external use*
ASL-1	No risk of catastrophe (e.g., chess-playing AI)	None	None
ASL-2 *Anthropic's current safety level*	Early indications of capabilities likely to cause catastrophe (e.g., unreliable bioweapon info)	Evaluate for ASL-3 warning signs; basic cybersecurity against opportunistic attackers	Misuse escalation procedures; vulnerability reporting; etc.
ASL-3 *Anthropic is currently preparing these measures*	Low-level autonomous capabilities or substantial increase in catastrophic misuse risk	Evaluate for ASL-4 warning signs; intermediate cybersecurity; compartmentalize training techniques and model hyperparameters	Strong misuse prevention measures; intensive expert red-teaming for all deployed modalities (e.g., API, fine-tuning)
ASL-4	*Anthropic to define capabilities and warning sign evaluations before training ASL-3 models*		

ASL-5+	

Table 1. Visualization of Anthropic's AI Safety Levels (ASL) system; adapted from Anthropic's RSPs (Anthropic 2023, p. 4)

As suggested above, Anthropic has not yet laid out concrete commitments for ASL-4. However, it has made some tentative suggestions for thresholds and response measures.

- Tentative capabilities thresholds for ASL-4 include:
  - AI models becoming the primary source of national security risk in a given domain;
  - AI models being able to autonomously replicate;
  - AI models being able to conduct autonomous AI *research*, which could greatly boost a malicious AI program.
- Tentative response measures include:
  - Security measures that exceed those of even the strongest current technology companies.
  - An "affirmative case" that models will not attempt to undermine safety measures or cause catastrophe.
  - The use of automated harm detection for all model use.
  - Employing external audits to verify the above measures.

Selection of DSIT best practices

Our review of Anthropic's RSP was conducted as a collaborative exercise between the four IAPS researchers who authored this paper. To narrow the scope of the exercise, we each independently selected six DSIT sub-recommendations that we thought were essential for RSPs to function: that is, they fulfilled the criterion of "Without this sub-recommendation being implemented, any RSP will fail to significantly reduce catastrophic risks." We then identified points of agreement and deliberated substantive disagreements. Here, we explore sub-recommendations that at least half of us thought were essential.

We ended up with eight such sub-recommendations, outlined in the below table. Throughout this piece, however, we group the third and fourth sub-recommendations (on risk-thresholds) together, meaning that in practice we analyze seven. The sub-recommendations are listed in the order that they appear in DSIT's guidance (UK DSIT, 2023).

DSIT sub-recommendation selected	Our reasoning

"**Develop rigorous risk assessment processes** for models."	Without good risk assessment techniques, the evaluator will not be able to determine which risk tier a given model belongs to, nor if any mitigations have reduced the level of risk sufficiently to allow continued development.
"Monitor systems [for risk] both **during development and after deployment.**"	Models may pose risks already during the development stage, for example, if leaked or stolen. New risks may emerge after deployment, for example, when a model is fine-tuned, connected to new tools, or used in unexpected ways.
"Describe and continually refine risk assessment results for each model ('**risk thresholds**') that would trigger particular risk-reducing actions."	Without risk thresholds, risk-reducing actions cannot be systematically specified in advance.
"**Operationalise** risk thresholds."	We cannot reliably tell if risk thresholds have been passed if we do not have a clearly demarcated boundary.
"At each risk threshold, proactively commit to **only proceed** with certain development or deployment steps **if specific mitigations are in place**."	Risk thresholds on their own do not increase safety, unless paired with corresponding risk-reducing actions.
"**Inform relevant government authorities** when a risk threshold has been met."	Government awareness will be essential to the mitigation of certain risks (e.g., catastrophic misuse).
"**Prepare to pause training runs or reduce access to deployed models**, if risk thresholds are reached without the committed risk mitigations being in place."	The stakes are high enough (particularly in higher ASL levels) that society cannot afford for companies to continue with development/deployment if risk mitigations are not in place.
"**Include verification mechanisms**, such that external actors can have increased confidence that responsible capability scaling policies are executed as intended."	We cannot reliably tell if risk thresholds have been passed, and corresponding risk-reducing actions implemented, without external verification.

How does Anthropic's RSP compare?

Below, we evaluate Anthropic's RSP against each of the DSIT sub-practices identified above.

Developing rigorous risk assessment processes for models

Anthropic's RSP focuses explicitly on catastrophic risks, "defined as large-scale devastation (for example, thousands of deaths or hundreds of billions of dollars in damage) that is directly caused by an AI model and wouldn't have occurred without it." We believe this focus on catastrophic risks to be appropriate for RSPs, as the RSP framework aims to identify and mitigate risks that originate specifically from scaling up model capabilities. While other harms such as bias and discrimination are also important, many of these already manifest in smaller models and are not purely a product of scaling.

Anthropic commits to employing a mixture of *forecasts* of future capabilities and *evaluations* ("evals") of present capabilities. Given the challenges of risk assessment for advanced AI systems, this combination of techniques provides a powerful "one-two punch": (1) forecasts of future capabilities are essential to taking timely action but are presently quite difficult, but (2) evaluations of present capabilities could provide a valuable stop-gap by making clear that if a model is displaying a dangerous capability now, risk-mitigation steps need to be taken.

On the other hand, evaluations of model capabilities may significantly underestimate the actual risks posed by models.

One problem here is that evaluations often only provide an *example* of dangerous capabilities elicited by the model, not the upper bound. Tweaks to evaluations can often lead to significant increases in the capabilities demonstrated by models. For example, as noted by Apollo AI research, a UK government partner, Bsharat et al. (2023) find that offering to tip an LLM $300,000 for a "better" response causes the model to demonstrate increased capabilities, while using more structured elicitation techniques such as chain-of-thought prompting can also lead to greatly improved capabilities (Wei et al, 2022). Apollo states that the science of eliciting maximal capabilities (rather than average capabilities) is only in its earliest stages.

Another problem potentially leading to the underestimation of model capability is that, just as current ML systems often face training incentives to make human evaluators think that they have a certain capability when in fact they do not, future ML systems may face incentives to deceive human evaluators into thinking that they do *not* have certain capabilities when in fact they *do* (see e.g., Hendrycks, Mazeika and Woodside, 2023; Park et al, 2023). Some researchers have termed that problem "deceptive alignment" (e.g., Carranza et al., 2023). Because of deceptive alignment, evaluations of model capabilities may not accurately reflect the

real risks posed by a given (future) system. To address this problem, companies and governments should fund the development of evaluations of deceptive alignment, such as those currently being pursued by Apollo.[8]

In some cases, problems such as the upper-bound issue and deceptive alignment may make formal risk-assessment extremely difficult, if techniques that deal with these problems are not available. For example, if and when their fourth (and currently highest) risk tier is breached, Anthropic commits to making an "affirmative case… that our models will not autonomously attempt to strategically undermine our safety measures or cause large-scale catastrophe." In the absence of deceptive alignment evaluations or good upper-bound evaluations, producing such a foolproof case may be close to impossible. For that reason, companies and governments should invest in targeted measures such as mechanistic interpretability research, evaluations of deceptive alignment, and other measures that could reduce the likelihood of under-estimating model capability.

Monitoring systems throughout development and deployment

Anthropic's RSP involves a certain degree of monitoring throughout both the development and the deployment phases.

Evaluations during the training and fine-tuning phases must take place both every three months and after every 4x jump in computing power.

Post-deployment, however, Anthropic's monitoring controls seem to be limited to automated detection "for attempts to cause harm, exfiltrate weights, or make changes to training runs." We think automated harm detection is a strong measure – it aligns with previous recommendations from IAPS that companies adopt judicious use of automated input-output monitoring to detect anomalous or harmful model use or behavior as a valuable AI risk mitigation tool (O'Brien, Ee, and Williams, 2023). However, Anthropic and other companies could more explicitly identify the use of post-deployment monitoring tools and mechanisms in RSPs. For example, in cybersecurity, Security Operations Centers (SOCs) are common among large companies that need to actively monitor systems 24/7 for cybersecurity threats and incidents.[9]

On a related note, while we think that automated harm detection throughout the product lifecycle would boost model safety and security, we think that detection and monitoring for

[8] If good deceptive alignment evals can be developed, the probability of deceptive alignment in a given case could be combined with the (independent) assessment of the likelihood that a given capability is present, to produce a more accurate uncertainty distribution over the likelihood that said capability is in fact present. We think that it is vital that model developers accurately communicate their uncertainty to policymakers.

[9] To ensure 24/7 coverage, large SOCs often adopt a globally distributed, "follow-the-sun" approach, where an office in an earlier time zone will hand off work to another in a later time zone.

high-risk systems should probably also involve human elements to ensure reliable oversight. For example, the EU AI Act now requires human oversight for high-risk models.

Describing, operationalizing and continually refining risk thresholds

Anthropic's RSP is based around the idea of "capability thresholds" which, once crossed, trigger a higher risk tier ("AI Safety Level" or ASL) involving correspondingly more intensive risk-reducing actions. Anthropic also notes that their risk thresholds for ASL-4 are provisional and will be revisited in light of further information. The idea of capability thresholds is a promising start, and Anthropic's ASL framework is a clever approach that borrows from risk tiers in the high-risk domain of biosecurity.

However, Anthropic's risk-thresholds for ASL-3 and ASL-4 are currently somewhat vague. For example, Anthropic's current ASL-3 misuse threshold is a "significant increase" in the risk of catastrophic misuse above baseline levels. Meanwhile, Anthropic's (more tentative) ASL-4 misuse threshold is that "AI models have become the primary source of risk in a major area (such as cyberattacks or biological weapons)."

The problem with such risk thresholds is that they are not specific enough to be verifiable – different experts may easily disagree over how to interpret either threshold. A more concrete approach to catastrophic risk management, commonly used in other industries where catastrophes might be plausible, is to set specific thresholds for "societal risk" (SR). SR captures the intuition that risk-tolerances for events involving multiple fatalities ("societal risks") should often be more stringent than risk-tolerances for isolated individual fatalities – for example, because the latter will often be more difficult to prevent, whereas a catastrophe involving multiple fatalities usually indicates a serious safety failure.

A non-exhaustive survey by Flamberg et al. (2016) provides some insight into the range of SR tolerances used across industries and countries. Based on that survey, *maximum* acceptable risk-tolerances for events involving 1000 or more deaths (Anthropic's definition of a catastrophic risk) have ranged from:

- An upper bound (= most tolerant) of 1 E-04 events per year.[10] This risk-tolerance was originally set by the UK Health and Safety Executive (HSE) in 1991, although it was later revised to an order of magnitude lower. The HSE sets standards in a variety of domains ranging from occupational health to nuclear power to biosecurity. The 1 E-04 threshold is still used by some organizations, for example, the International Maritime Organization, for certain purposes.

[10] Equivalent to a 0.01% *probability* of such an event per year, or a 1% probability per century.

- A lower bound (= least tolerant) of 1 E-10 events per year.[11] That threshold is used by the Czech Republic for hazardous facilities in multiple industries, although it may be too stringent to be realistic for many such facilities (Flamberg et al, 2016).

Standards-setting bodies also often set "broadly acceptable" SR tolerances, generally two orders of magnitude lower than their maximum risk tolerances (Flamberg et al, 2016).

We tentatively suggest that AI companies' thresholds for ASL-4 and ASL-3 (or equivalent) should be set in line with the above-outlined "maximum" and "broadly acceptable" SR ranges, respectively. Of course, very low probabilities can be hard to measure, but we suggest that for higher-magnitude catastrophic risks (or catastrophic risks where the upper bound of impact may be difficult to determine), a proportionally lower tolerance in the range should be adopted. (For example, there should be a proportionally lower tolerance for events involving 10,000, 100,000, and 1 million deaths respectively). That intuition is often captured via the use of F-N curves (Flamberg et al, 2016). **Because some risks may resist quantification even via subjective judgment, we suggest that more qualitative risk thresholds be employed in tandem with the above approach.**

The risk thresholds that we suggest are probably lower than Anthropic's current risk thresholds. For example, the wording ("primary source of risk") in Anthropic's above-outlined ASL-4 misuse threshold implies that catastrophic misuse risk from AI in a given domain (e.g., biosecurity) should be equal to the baseline non-AI risk in that domain. However, research from Piers Millett and Andrew Snyder-Beattie (2017) estimates that the annualized risk of a (non-AI) bioweapons disaster killing upwards of 1,000 people is above 1%. That is, 1 E-02 catastrophic events per year on average – two orders of magnitude above the upper bound SR tolerance outlined above.

Another problem with Anthropic's current risk thresholds (at least for ASL-4), which using an SR-style threshold would solve, is that they rely on a relative rather than an absolute definition of risk:

- *Relative risk* focuses on the *proportion* by which a given risk has changed from the baseline.
- *Absolute risk* focuses on the *absolute value* by which a given risk has changed from the baseline.

Anthropic's ASL-4 misuse risk threshold, at least, employs a kind of relative risk – ASL-4 will kick in once AI risk is *equal to* the non-AI baseline, a proportional threshold. (It is unclear whether Anthropic's other risk-thresholds are absolute or relative).

[11] Equivalent to a 1 E-08 % *probability* of such an event per year, or a 1 E-6 % probability per century.

We recommend that Anthropic and other AI companies solely adopt a framework of absolute risk (for example, the SR framework outlined above), rather than relative risk. That follows common advice in other domains (e.g., healthcare: see Jiroutek and Turner, 2019).

One benefit of using an absolute risk threshold is that the baseline risk may be extremely difficult to determine, with different studies suggesting different baseline risks. If the risk threshold is relative, i.e., proportional to the baseline risk, there is a danger that less scrupulous companies will simply pick and choose baseline risks to suit their interests (for example, claiming that the baseline risk is very high and that the risk from their model does not match it, even if the model threatens a high absolute level of harm).

A closely-related problem with using relative rather than absolute risk is that the former involves a moving target. For instance:

- If non-AI biological misuse risk increased to 10% per year due to biotechnology advancements, AI-biological misuse risk would have to increase to 10% per annum before ASL-4 was triggered.
- Similarly, if non-AI biological misuse risk was to fall to a sufficient level *below* AI misuse risk (e.g., due to improvements in management of biological misuse risk), ASL-4 could be triggered despite AI misuse risk not having increased at all.

Anthropic's risk thresholds are also possibly not specific enough laterally, as well as vertically. By that we mean that there is insufficient breakdown of different risk *types*, as well as risk *levels.* For instance, a "significant" increase in cyber misuse risk could look very different from a "significant" increase in biological misuse risk, and would almost certainly entail different risk-reducing actions.

Note that while Anthropic does not currently specify risk thresholds for these categories, it does commit to an intended plan of work to do so, stating "for each domain of interest (CBRN and cyber risks)" it will define precise threat models, specify capability improvements that would enable those threats, and design model evaluations for those capabilities. It also does in practice assess risk by domain - for instance, its model card for Claude 3 includes separate evaluations of three capabilities: Autonomous Replication and Adaptations, Biological, and Cyber. Each domain has specific ASL-3 boundaries (e.g. the autonomous replication and adaptations boundary is 'the model passing 50% or more of the tasks described below with at least 10% pass rate (i.e. succeeding at one in ten attempts'). However, separate boundaries are currently not specified with regard to societal risk / outcomes, and only three risk domains are covered. This makes it difficult to assess the societal risk posed by the model, and to know if the risk assessment appropriately covers all risks that could trigger ASL-3.

Companies should therefore set distinct SR thresholds, with distinct corresponding measures, for a more granular subset of risks – for example, not just "misuse" but "cyber misuse" as distinct from "biological misuse." To better taxonomize some of these

potential risks, companies could use a "techniques, tactics, and procedures" framework, as is done in cybersecurity with the MITRE ATT&CK framework (Ee et al., 2023).

This process will be iterative, and there should not be an expectation that Anthropic, or other companies, will immediately be able to identify a collectively exhaustive taxonomy for such risks.

Instead, companies could commit to a work plan to outline such risks by a set deadline, potentially collaborating in forums such as the Frontier Model Forum or the NIST AI Safety Consortium. The work plan could include statements such as:

- "We identify A, B, C domains as particularly important for risk of catastrophic misuse, and by X date, we will publish our best assessment of what the risks are in A, B, and C."
- "Additionally we consider D, E, F domains to be among those that warrant further scrutiny, and we will continue to investigate them and will establish a work plan at a future date."

In any case, all of our suggestions in the above section are tentative, and standardized risk-thresholds for RSPs should ultimately be set by a government body, such as UK DSIT or the National Institute of Standards and Technology (NIST), or a corporate body, such as the Frontier Model Forum (FMF). Clearly operationalized thresholds will make it easier to reach agreement between different evaluators as to whether the risk threshold has been breached. Otherwise, even if Anthropic have strong internal risk thresholds that are not published as part of their RSP, AI racing dynamics mean that said thresholds may be of little use unless adopted by other companies.

Committing to only proceed with development/deployment if certain mitigations are in place

One strength of Anthropic's RSP is that it commits the company to undertaking certain mitigations once a risk threshold is crossed, and pausing (see below) until said mitigations are in place. For example, for systems that reach ASL-3, Anthropic commits to undertaking "strong misuse prevention measures, including internal usage controls, automated detection, a vulnerability disclosure process, and maximum jailbreak response times."

Another strength of Anthropic's RSP is that it recognizes that risk-mitigation should not simply be procedural (e.g., "undertake evaluations"), but should also specify targets for risk-reduction, such that risk-reducing actions must continue until the risk is below said target (e.g., "evaluations should come up negative for dangerous capabilities XYZ"). For example, Anthropic's RSP specifies that for models that pass the ASL-4 risk threshold, Anthropic will need to see affirmative evidence of safety (rather than merely have evaluations fail to show signs

of danger) – even if such affirmative evidence is currently not possible to produce. We believe this sets an important precedent that other companies should follow.

However, Anthropic's risk-reduction measures lack granularity in several places, again in part because Anthropic's risk thresholds themselves are not yet operationalized in detail. This is especially true for risk-reduction measures intended to address misuse, which are difficult to assess given the lack of public information on the risk levels and subtypes that would trigger them. Without knowing what the concrete level or type of risk is, we cannot assess whether the proposed measures are adequate to reduce said risk.

Informing relevant government authorities once a given risk threshold is crossed

Anthropic's RSP currently lacks detail about sharing information with government authorities. The only mention of any such sharing is as follows:

> "**In a situation of extreme emergency**, such as when a clearly bad actor (such as a rogue state) is scaling in so reckless a manner that it is likely to lead to lead to imminent global catastrophe if not stopped (and where AI itself is helpful in such defense), **we could envisage a substantial loosening of [RSP] restrictions as an emergency response. Such action would only be taken in consultation with governmental authorities.**"

This wording implies that Anthropic will only alert governments under a specific, narrow instance of potential catastrophic threats, and only in order to discuss the loosening of RSP restrictions so that governments can more powerfully respond to said threats. For other scenarios and ASL levels, Anthropic has not made any commitments to government reporting or collaboration in their RSP.

We believe that Anthropic and other AI companies should commit to sharing *all* risk-assessment and evaluation results above a certain threshold with relevant governments agencies. That threshold should probably be lower than an "imminent" chance of a global catastrophe, for example, the tentative ASL-3 and ASL-4 thresholds that we outline above.

Preparing to pause training runs/reduce model access

The "procedural commitments" in Anthropic's RSP specifically identify the need to "proactively plan for a pause in scaling if one proves necessary… to implement security or other measures required to support safe training and deployment." This commitment is valuable because the costs of delaying model development could be substantial, and if companies are not appropriately prepared to implement a pause, that may disincentivize their willingness to do so.

AI companies should also consider how to coordinate with each other where delays are necessary. For example, if a key risk threshold is passed by one company, that company may need to inform other companies, and these companies may need to pause their own training runs to assess risk thresholds. Failure to coordinate could reduce the effectiveness of individual companies pausing: e.g., customers may switch from a more responsible company that has paused development to a less responsible company. Upcoming work from IAPS on coordinated disclosures of dangerous capabilities will explore this process.

AI companies, and governments, should also consider what processes or infrastructure should be put in place to ensure that pauses in model development can be implemented smoothly. For example, service level agreements (SLAs) with downstream providers could be drafted to account for potential emergency pauses (O'Brien, Ee, and Williams, 2023).

Including external verification mechanisms

Anthropic's RSP does not consider external verification mechanisms in much detail. Most notably, it does not mention external verification mechanisms for any of its evaluations until systems reach the ASL-4 threshold (when the threat of an AI-related global catastrophe would be extremely high compared to current levels). It simply states:

> "Due to the large potential negative externalities of operating an ASL-4 lab, verifiability of the above measures [security, safety research, evals, automated harm detection] should be supported by external audits."

We suggest external audits should instead start to occur at ASL-3, given that ASL-3 entails a "substantial" increase in catastrophic misuse risk. External validation should be adopted in high-stakes scenarios that could result in human deaths, as is done to improve process and product reliability in other safety-critical industries. For example, in the aviation industry, final manufacturers typically conduct inspections of their suppliers, and government authorities in turn are supposed to inspect final manufacturers; failures in this series of verification mechanisms led to blowout of a panel in an Alaska Airlines plane that could have been catastrophic (Schwenk et al., 2024; Chokshi and Walker, 2024).

The ASL-3 "Evaluations for Misuse Risks" appendix does mention collaboration with external experts on *developing* evals for biological, cyber, and general CBRN risks, but it does not commit to getting evaluation *results* verified by external auditors. We think that the latter would be a valuable next step.

Conclusion

Overall, Anthropic's RSP demonstrates a number of strengths that other AI companies should seek to copy. It also demonstrates a number of limitations that either Anthropic or government authorities should seek to patch, and from which other companies should learn when developing their own versions of RSPs. Based on those strengths and limitations, our report has suggested some tentative recommendations to guide the future actions of such organizations.