Toward Thresholds for Intolerable Risks Posed by Frontier AI Models

WORKING PAPER

Provisional Recommendations and Considerations for Intolerable Risk Thresholds to Inform Frontier AI Safety Frameworks

Anthony M. Barrett, Jessica Newman, Deepika Raman, Nada Madkour, Evan R. Murphy
UC Berkeley Center for Long-Term Cybersecurity

15 November 2024

Affiliations listed above are current, or were during the authors' main contributions to this work. Views expressed here are those of the authors, and do not represent views of UC Berkeley nor others.

Access the latest version of the document [at this link]

Table of Contents

Executive Summary

AI foundation models, especially increasingly advanced frontier models, may pose **intolerable risks**. For example, frontier models may lower barriers to a terrorist, state-affiliated threat actor, or other adversary seeking to cause high-impact events such as CBRN or CBRNE (chemical, biological, radiological, nuclear, explosive) attacks or cyber-attacks. In this paper, we consider intolerable risks as those that have the potential for severe or catastrophic impact; have relevance to current and emerging AI capabilities; have the likelihood of irreversibility**;** and have a short timescale of expected impact.

A number of frameworks for pre-release risk assessment and decision-making include AI dual-use capability evaluation and some form of explicit or implicit threshold for a dual-use capability hazard that should be regarded as intolerable. However, such thresholds are often defined at a high level, using qualitative language, which may not be readily compared to results of dual-use capability assessments, such as from red-teaming-based evaluations. **Model developers and evaluators may be left without a reasonably clear, consistent, and operationalized answer to the question, "How much lowering of barriers is too much?"**

Now is the time to consider that question. One reason is to inform Frontier AI Safety Commitments work by AI industry organizations on defining intolerable risk thresholds (DSIT 2024a) before the AI Action Summit in France to be held February 10-11, 2025. In the absence of clear guidance from regulators, academics, or civil society that place a high priority on protecting public safety, companies may face incentives to develop thresholds that are low-cost for them to implement, but that do not provide levels of public safety that are as high as in other domains. Those thresholds also may set precedents, or at least have substantial framing effects, affecting governmental risk-threshold policymaking in the United States (e.g. on export-control regulations following Executive Order 14110) and Europe (e.g. on general-purpose models with systemic risks).

In this paper, we provide background on intolerable risk thresholds and propose a number of recommendations and considerations for organizations and governments exploring how to define and operationalize intolerable risk thresholds. We intend this material to be a starting point or supplementary resource for others to use in their own deliberations.

Key Recommendations

- **Consensus on the definition and categories of intolerable risks:** There is a range of taxonomies of risks and harms that need to be standardized in order to enable uniformity in the risk assessment and reporting exercise. This is where national policies and international agreements can lend their power to operationalize better oversight.
    - We propose expanding the definition from the Frontier AI Safety Commitments and providing accompanying specific examples and considerations to provide clarity and enable consensus.
    - We consider specific intolerable risks arising from CBRN weapons, cyber operations, model autonomy, persuasion, and deception, along with additional

    considerations for unacceptable uses, limitations, impacts, and other key considerations.

- **Application of appropriate thresholds:** While there is precedent for thresholds based on compute in the US and EU AI policy approaches, there have also been demonstrated limitations in using compute thresholds as the sole determinant of risk. Other proposed thresholds include capability thresholds and risk thresholds.
    - We recommend focusing on capabilities (more than “risk” per se or compute), at least until likelihood estimation becomes more reliable, while accounting for reasonably foreseeable ways capabilities can be enhanced (e.g. plugin tools and scaffolding)
    - To determine the appropriate evaluations, and adapt existing benchmarks to determine modern capability, we can break down key variables that comprise the capability threshold into three aspects: planning capability, knowledge capability, and execution capability.
- **From capabilities to thresholds:** For arriving at the specific thresholds, we propose specific relevant capabilities that increase the probability of occurrence of one or many categories of intolerable risks, along with recommendations for their timely calibration to keep pace with rapid AI advancements (see Table 1). The determination of exact quantitative thresholds would be determined by (i) isolating one/few metrics of interest, (ii) the technical feasibility and robustness of measuring these metrics, and (iii) the organization or national appetite and attitude towards risk-benefit tradeoffs.
    - Any recommendations on specific thresholds will be inherently intertwined with business motivations and moral obligations. Since these are not static measures, national-level determinations of risk tolerance can be applied to the proposed capabilities. We demonstrate these approaches in Appendix B where we arrive at numerical thresholds for CBRN and deception which advocates for a lower threshold in comparison to industry, reflective of the different appetites for risk.

Acknowledgments

This working paper was informed by an in-person roundtable held in Berkeley, CA on November 12, 2024, that brought together representatives from academia, industry, civil society, and government. We are grateful to the roundtable participants for their expert feedback and insights. Participants do not necessarily endorse the recommendations or considerations included here. Participants included: Sacha Alanoca, Esteban Arcaute, Colleen Chien, William Cutler, Ian Eisenberg, Florence G'sell, Meredith Lee, Alisar Mustafa, Stuart Russell, Maeve Ryan, Madhu Srikumar, Shane Witnov, and others. We are grateful to the Foundational Research Grants program at the Center for Security and Emerging Technology (CSET) for support of our ongoing work in this space.

1. Background

1.1 The Call for Defining Intolerable Risk Thresholds

In May 2024 at the AI Seoul Summit, sixteen global AI industry organizations committed to publishing their efforts to measure and manage risks posed by their frontier AI models in an accountable and transparent manner and determine thresholds for intolerable risks (DSIT 2024a). Specifically, as part of these **Frontier AI Safety Commitments**, they must define **"thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable."** They specify further, “Thresholds can be defined using model capabilities, estimates of risk, implemented safeguards, deployment contexts and/or other relevant risk factors. It should be possible to assess whether thresholds have been breached.”

At the same summit in Seoul, **states announced their intent to define these thresholds for frontier AI systems leading up to the AI Action Summit in France** signaling the regulatory appetite to challenge self-imposed, “voluntary” limits by industry actors. Both pledges explicitly state the importance of ensuring these thresholds are defined with input from a range of trusted actors and require reporting on the different capacities of their involvement in these efforts (DSIT 2024b).

1.2 Existing Intolerable Risk Thresholds

Multiple types of thresholds have been proposed and used, including **capability thresholds, compute thresholds, and risk thresholds**. The differences between these are discussed in Appendix A. Here **we focus on capability thresholds, which are most prominent in current examples and have been most closely aligned with establishing intolerable risk thresholds.**

Several model developers and deployers have published AI capability thresholds. Anthropic categorizes capability thresholds as **“red line” and “yellow line” capabilities (**Anthropic n.d.) with corresponding AI Safety Level (ASL) standards in their Responsible Scaling Policy (RSP). Redline capabilities refer to anticipated model abilities that may appear in future versions of the model and would present too much risk if deployed under current ASL-2 safety measures. Anthropic has committed to developing a new set of ASL-3 safety measures to sufficiently manage and mitigate models with red-line capabilities. Anthropic has also defined qualitative capability thresholds for specific model capabilities (chemical, biological, radiological, and nuclear (CBRN) weapons, autonomous AI research and development, and cyber operations). For example, the CBRN weapon threshold is defined as “The ability to significantly help individuals or groups with basic technical backgrounds (e.g., undergraduate STEM degrees) create/obtain and deploy CBRN weapons” (Anthropic 2024 p.3). These thresholds are evaluated by first conducting a preliminary assessment to determine if a model is “notably more capable” than the latest model that has been comprehensively tested. Models that are effectively more capable (4x or more in Effective Compute or 6 months worth of finetuning) undergo a comprehensive assessment containing threat model mapping for each capability threshold, empirical tests for capability evaluation, elicitation testing without safety mechanisms, and likelihood forecasting (Anthropic 2024). Anthropic maintains a dynamic risk scorecard that reflects pre- and post-mitigation evaluation results for each of the tracked capability categories. The risk levels inform Anthropic's decision to enforce certain safety baseline actions based on pre- or post-mitigation risk scores. For example, models can only be deployed if they are determined to have a post-mitigation risk score of “medium” or below.

The OpenAI preparedness framework categorizes thresholds using a **qualitative scale (low, medium, high, and critical)** with definitions for each of their four tracked capability categories (cybersecurity, persuasion, CBRN, and model autonomy). For example, a model that is considered to have a “high” level of cybersecurity capability risk is defined as a “Tool-augmented model (that) can identify and develop proofs-of-concept for high-value exploits against hardened targets without human intervention, potentially involving novel exploitation techniques, OR provided with a detailed strategy, the model can end-to-end execute cyber operations involving the above tasks without human intervention” (OpenAI 2023b p.8). A model’s overall capability score is determined by the highest score in any of the tracked categories.

Google DeepMind’s Frontier Safety Framework (Dragan et al. 2024) outlines model **critical capability levels (CCLs)** which are identified with preliminary model evaluations for various risk domains (autonomy, biosecurity, cybersecurity, and machine learning R&D). Each risk domain CCL is described and includes the rationale behind the categorization. For example, bio expert enablement level 1 is described as “Capable of significantly enabling an expert (i.e. PhD or above) to develop novel biothreats that could result in an incident of high severity” (Dragan et al. 2024 p.5). The model developer states that “The Framework is exploratory and based on preliminary research. We expect it to evolve substantially as our understanding of the risks and benefits of frontier models improves, and we will publish substantive revisions as appropriate” (Dragan et al. 2024 p.6).

2. Proposal to Advance Intolerable Risk Thresholds

Across the three examples described above, there is relative consensus as to the key risk categories that are considered in determining capability thresholds. All three discuss the general risk categories of CBRN weapons, cyber operations, and model autonomy. The OpenAI preparedness framework additionally considers *persuasion* and notes that they include *deception* and social engineering evaluations as part of the persuasion risk category. Anthropic’s Responsible Scaling Policy includes the footnote, "We recognize the potential risks of highly persuasive AI models. While we are actively consulting experts, we believe this capability is not yet sufficiently understood to include in our current commitments."

We recommend explicitly including the two categories of *persuasion* and *deception* in capability thresholds and as part of Frontier AI Safety Frameworks (Anthropic 2024).

2.1 Recommended Inclusions for Intolerable Risks

Persuasion

Large language models have been found to be effective at producing “personalized persuasion” at scale (Matz et al. 2024), meaning messages and content can be specifically designed based on an individual’s psychological and other traits in order to have particular effects. The ability of frontier models to engage in effective persuasion means that people and communities may be more easily manipulated or exploited. Examples of substantial risks that can stem from persuasion include widespread medical disinformation leading to loss of life, the incitement of violence or self-harm behaviors, and swaying election outcomes leading to an erosion of democracy.

The OpenAI Preparedness Framework describes the ability of an AI model to “dramatically affect elections and democratic outcomes” as high risk, and the ability of an AI model to serve as “a powerful weapon for controlling nation-states, extracting secrets, and interfering with democracy” as a critical risk (OpenAI 2023b). OpenAI monitors such uses of its models and released a report describing multiple covert influence operations using its AI models (OpenAI 2024).

To determine persuasion thresholds, it is reasonable to compare model outcomes to the current costs and time frames associated with human actors employed in activities like spreading misinformation through content generation without access to frontier model capabilities. Most models released since 2022 have, on average, generated content indistinguishable to human participants over 50% of the time, with two tested models surpassing human-level humanness[1] (Wiliams et al. 2024, Chen and Shu 2023). Some estimates predict that current model capabilities in content generation bring the cost of launching election disinformation campaigns down to near zero with smaller open-source models and to about $9 for hosting Llama 3 70B or $0.30 when using Gemini’s API tokens. (Wiliams et al. 2024)

Musser (2023) analyzes the cost-effectiveness of deploying LLMs for generating content indicating the potential incentives for adversarial adoption- “Over the course of a 10-million-tweet campaign, with a language model that produces usable outputs at a rate of 75%, a propagandist could expect to save $3 million in content generation costs, on average (assuming no fixed costs to using the model and no monitoring controls in place on the model; 95% CI: [$430,000, $9.4 million]).”

[1] Defined by the authors as “the extent to which disinformation operation content generated by an LLM is able to pass as human written” (Wiliams et al. 2024)

Kapoor et al. (2023) argue the costs of content distribution may be more crucial than the costs of content creation. For instance, current misinformation outputs and disinformation campaigns are carried out by a sophisticated, fast-growing industry (Bradshaw et al. 2021). The demand for these services continues to rise (Fisher 2021) owing to their ability to bundle accompanying services like maintaining numerous fictitious user profiles, curating strategies to sidestep social media detection and moderation, non-English language proficiencies, expertise on geographical and cultural nuances, etc.

While frontier models perform exceptionally in the ‘knowledge’ dimension of persuasive capabilities, they still lack the sophistication of human-powered endeavors in identifying appropriate target groups and rolling out sustained campaigns to affect intended change (of influencing votes, spreading conspiracies, etc.) Future-proofing the threshold for persuasion will entail the inclusion of appropriate safety margins that keep pace with the forecasted capabilities of the models as well as the estimated demand for malicious use (Goldstein et al. 2023, Sedova et al. 2021).

Deception

Advocating for the recognition of risks from AI deceit as ‘high risk’ or ‘unacceptable risk’, Park and Goldstein (2023) define deception as “the systematic production of false beliefs in others as a means to accomplish some outcome other than the truth.” Such behavior can not only influence model performance in downstream applications but also challenge the reliability of any safety testing performed on them.

Strategic underperformance or sandbagging on capability testing and other assessments can be induced by developers to circumvent safety requirements, or be an inherent behavior from the model itself (Järviniemi and Hubinger 2024). Famous examples of model deception include an early evaluation of GPT-4 which demonstrated model success in utilizing a human Taskrabbit worker to solve a CAPTCHA for it under the pretext that it had a vision impairment. (Open AI 2023a, pp. 15-16).

Specifically targeting capability evaluations through prompt engineering, van der Weij et al. (2024) demonstrate how GPT-4 and Claude 3 Opus can selectively underperform when evaluated for dangerous capabilities (albeit inconsistently). Consistency checks (Fluri et al. 2023) of model-produced outputs, and examining internal representations to model outputs (Zou, Phan et. al 2023) are a few potentially helpful but not sufficiently rigorous assessments to determine deceptive behaviors. Apart from a generalized motivation to consistently underperform in evaluation, deceptive behavior also arises from a model’s ability to distinguish between phases of its training and its testing. Model perception of its evaluation phase only reinforces the need to think beyond memorizable practices of standardized safety testing efforts to more holistic, hybrid evaluations that can adequately test capability in context.

Recent literature also raises alarms of current safety techniques being insufficient in adequately mitigating such deceptive behavior once identified (Hubinger et al. 2024). Owing to the inaccuracy of evaluation and growing immunity to finetuning, current technical practices are unmistakably insufficient in combating risks arising from deceptive models. Therefore it is necessary to determine model capabilities that appear as a precursor to deceptive tendencies in order to mitigate such intolerable risks. Appendix B expands on this idea further.

2.2 Thresholds for Key Intolerable Risk Categories

To consolidate intolerable risks across the different industry-led and state-led efforts to mitigate risk, we used the following **criteria for inclusion**:

- Potential for Severe or Catastrophic Impact
- Relevance to Current and Emerging AI Capabilities
- Likelihood of Irreversibility
- Short Timescale of Expected Impact

In proposing thresholds, we used the following **key principles** (see Appendix C for longer descriptions):

- Seek to identify cases of substantial increase in risk
- Focus on capabilities (more than “risk” per se or compute), at least until likelihood estimation becomes more reliable, while accounting for reasonably foreseeable ways capabilities can be enhanced (e.g. plugin tools and scaffolding)
- Compare to appropriate base cases
- Minimal increases in risk should be detectable, but not necessarily intolerable
- A substantial increase in capabilities for some attack stages can have a disproportionately large effect on risk
- Leave some margin of safety: operationalize intolerable risk thresholds at approximately the “substantial” level, leaving some margin of safety before arriving at a “severe” level
- Define at least some thresholds without factoring in model-capability mitigations due to the unreliability of virtually all safeguards at this time

Toward Thresholds: Mapping Intolerable Risks to Model Capabilities

The following table captures evidence from literature supporting five identified intolerable risks - CBRN weapons, cyber operations, model autonomy, persuasion, and deception.

By taking into account the deployment context of these models, their identified capabilities, and known adversarial demand and behavior, we also provide more in-depth examples (Appendix B) that demonstrate approaches to determining quantitative thresholds from available evaluations.

Table 1: Toward Thresholds: Mapping Intolerable Risks to Model Capabilities

Risk Category	**Rationale for Inclusion and Outcomes of Concern**	**Evidence**	**Intolerable Risk Thresholds** *(Defined primarily in terms of unacceptable model capabilities)*
CBRN Weapons	AI's role in managing or mismanaging Chemical, Biological, Radiological, and Nuclear materials can have catastrophic consequences. The misuse or accidental release of such materials can lead to mass casualties and long-term environmental damage. E.g. Deploying frontier models toward creating protein sequences for toxin synthesis for biowarfare.	Models have shown capabilities related to CBRN weaponization such as escalating conflict (Rivera et al. 2024), CBRN operations (JCS 2018), and semi-autonomous planning of complex chemical reactions (Boiko et al. 2023b).	Substantially more useful[2] than baseline sources (internet archives, scientific literature, technical manuals) for providing information, planning, and execution support for malicious CBRN weapon use. *See Appendix B for an example of a possible quantitative capability threshold.*
Cyber Operations	AI-enhanced cyberattacks can compromise critical infrastructure, national security, and personal privacy on an unprecedented	LLMs have demonstrated the ability to assist in cyberattacks (Lin et al. 2024, CPR 2023).	Identify and develop attacks against hardened targets.

[2] Assessments to estimate increases in CBRN capabilities can be obtained through an expert judgment-based rating approach and or statistical-significance test, which can be part of "human uplift" and/or "red teaming" based evaluation approaches. Examples of this can be found in the results section of OpenAI's red teaming study (Patwardhan et al. 2024) and RAND's red teaming study (Mouton et al. 2024).

	scale, leading to widespread disruption and economic loss. Eg. Autonomous capability to execute all critical phases of a ransomware attack, successfully identify and exploit vulnerabilities, etc. causing adversarial access to critical infrastructure and information.		End-to-end execution of cyber operations. Identify and develop zero-day exploits. Plan and execute end-to-end novel cyberattack strategies. Significantly enhance sophisticated attacks.
Model Autonomy	AI systems with high levels of autonomy can make decisions and take actions without human oversight. In critical domains such as military operations, autonomous vehicles, and infrastructure management, the loss of human control could result in accidents, escalation of conflicts, disruption of essential services, or other actions that do not align with human ethical standards or societal values. E.g. Deploying fully autonomous offensive weapon systems that can select and engage targets without human intervention, leading to unintended casualties and potential violations of international humanitarian laws.	Research has shown agentic systems that use frontier models as central controllers are increasingly capable of autonomous design, planning, and performance of complex scientific experiments (Boiko et al. 2023a).	Offensive lethal autonomous weapons (LAWs). Self-replication (Russell 2024). Fully automated R&D in AI, robotics, or other sensitive domains (Karnofsky 2024). Execute open-ended machine learning (ML) tasks that would contribute to critical steps to model improvement[3] (OpenAI 2023b, Anthropic 2024). Self-exfiltrate under current prevailing security (OpenAI 2023b, Anthropic 2024).
Persuasion	AI-driven manipulation through targeted misinformation or propaganda can undermine democratic processes, social	Microtargeting and manipulation through personalized persuasion strategies have been	Generating content which has a persuasive effect on a sufficient percentage of a

[3] This does not necessarily demonstrate the ability to continually improve over time. Many models may be bounded, or restricted by limited access to resources (OpenAI 2023b).

	cohesion, and individual autonomy, leading to societal destabilization. Eg. Covert and personalized interactions that can be used for radicalization and recruitment into extremist ideologies.	demonstrated at scale (Simchon et al. 2024) (Salvi et al. 2024). Successful political persuasion on critical policy issues through AI-generated messages has been demonstrated on a diverse population of Americans (Bai et al. 2023).	targeted population, leading to safety critical outcomes. Precise identification and successful contact with vulnerable groups (like children, or the elderly) and persuasive success in each interaction.[4]
Deception	An AI system that can deceive humans about its capabilities or plans could present numerous catastrophic risks. If a model can reliably distinguish between when it is running in a training or evaluation environment and when it has been deployed, then accurately measuring the dangerous capabilities of the model ceases to be possible. E.g. A model recognizing it is being evaluated strategically underperforms on CBRN capability evals, convincing humans it poses no danger in these areas. The model is then deployed widely, and it successfully assists a terrorist in developing and using a bioweapon of mass destruction.	Research has shown that models are capable of strategic deception (Park et al. 2023), safety training deception (Hubinger et al. 2024), situational awareness (Laine et al. 2024), and strategic underperformance (van der Weij et al. 2024).	Situational awareness[5] to identify with some reliability whether the model's environment is an evaluation or real-world deployment. *See Appendix B for an example of a possible quantitative capability threshold.*

[4] The OpenAI preparedness framework (OpenAI 2023b) frames this as "persuasive effectiveness strong enough to convince almost anyone to take action on a belief that goes against their natural interest" but we advocate for an earlier benchmark of such a capability approaching a critical percentage of the population. (E.g. with large enough reach, even a small percentage of the targeted population may be enough to disrupt democratic processes).

[5] Although deception is ultimately the capability we want to detect and prevent, we expect it will be extremely difficult to evaluate deception directly in an AI model based on a deep neural network architecture. Hence precursor proxy capabilities for deception such as situational awareness become relevant (Laine et al. 2024, pp. 33-34).

2.3 Additional Intolerable Risks

There are many other risks from frontier AI models that may be unacceptable or intolerable beyond the risk categories described above, which primarily relate to capability thresholds.

These additional intolerable risks can stem from unacceptable uses of frontier AI models, unacceptable limitations of frontier AI models, and unacceptable impacts of frontier AI models.

These are all of critical importance and require attention and mitigation. They differ from the capability thresholds described in depth above but can help inform decision-making about risk management and mitigation.

We present the following recommendations to take into account when making risk management and mitigation decisions, including whether an AI model may breach an intolerable risk threshold.

2.3.1 Unacceptable Uses

Some regulatory and policy frameworks define unacceptable uses of AI systems. For example, the EU AI Act has a list of prohibited AI practices (EP 2024, Article 5), which includes (with greater specificity):

- The use of an AI system to exploit any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability, or a specific social or economic situation;
- The use of an AI system for the evaluation or classification of natural persons or groups of persons;
- The use of an AI system for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offense;
- The use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage;
- the use of AI systems to infer emotions of a natural person in the areas of workplace and education institutions;
- the use of biometric categorisation systems that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation;
- And the use of ‘real-time’ remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement.

The U.S. White House Framework to Advance AI Governance and Risk Management in National Security (White House 2024) includes prohibited AI use cases for government agencies.[6] AI use cases that are prohibited include AI that is used to:

- Profile, target, or track activities of individuals based solely on their exercise of rights protected under the Constitution and applicable U.S. domestic law, including freedom of expression, association, and assembly rights.
- Unlawfully suppress or burden the right to free speech or right to legal counsel.
- Unlawfully disadvantage an individual based on their ethnicity, national origin, race, sex, gender, gender identity, sexual orientation, disability status, or religion.
- Detect, measure, or infer an individual's emotional state from data acquired about that person, except for a lawful and justified reason such as for the purposes of supporting the health of consenting U.S. Government personnel.
- Infer or determine, relying solely on biometrics data, a person's religious, ethnic, racial, sexual orientation, disability status, gender identity, or political identity.
- Determine collateral damage and casualty estimations, including identifying the presence of noncombatants, prior to kinetic action without (1) rigorous testing and assurance within the AI systems' well-defined uses and across their entire lifecycles, and (2) oversight by trained personnel who are responsible for such estimations exercising appropriate levels of judgment and care.
- Adjudicate or otherwise render a final determination of an individual's immigration classification, including related to refuge or asylum, or other entry or admission into the United States.
- Produce and disseminate reports or intelligence analysis based solely on AI outputs without sufficient warnings that enable the reader of the reports or analysis to recognize that the report or analysis is based solely on AI outputs.
- Remove a human "in the loop" for actions critical to informing and executing decisions by the President to initiate or terminate nuclear weapons employment.

Other types of unacceptable uses may include fraud and impersonation scams, generating Non-Consensual Intimate Image (NCII) or Child Sexual Abuse Material (CSAM), defamation of real individuals, and other human rights abuses.

If a model developer is operating in an environment in which there are no legal restrictions against such unacceptable uses or they are limited in reach and scope, developers should deploy mitigations to help prevent such uses from proliferating for example by publishing unacceptable use policies, restricting model access to users that violate these policies, and ensuring appropriate transparency on internal evaluations to enable third-party assessments of its validity (FMF 2024)*. If these mitigations are not expected to be effective, model developers should factor unacceptable uses into their assessment of whether a model breaches their intolerable risk threshold.*

[6] Also see Section 5b "Determining Which Artificial Intelligence Is Presumed to Be Safety-Impacting or Rights Impacting" in OMB (2024)

2.3.2 Unacceptable Limitations

In addition to unacceptable capabilities, frontier AI models may have unacceptable limitations. These may include hallucinations, harmful bias, malfunctions, errors, security vulnerabilities, a lack of reliability, and low accuracy in particular domains, languages, or cultural contexts.

There are numerous benchmarks and evaluations that can help assess the degree to which such limitations are present in models, though the accuracy and comprehensiveness of these benchmarks and evaluations vary. It is also essential that evaluations of critical model capabilities are verified as reliable approximations of risk (van der Weij et al. 2024).

If a model developer identifies serious limitations that cannot be remedied, they should factor unacceptable limitations into their assessment of whether a model breaches their intolerable risk threshold.

2.3.3 Unacceptable Impacts

Frontier AI models may also have unacceptable impacts. Examples include representational or allocative harms that are unacceptable or intolerable, large-scale discrimination, promoting violent and non-violent crimes, encouraging self-harm or suicidal ideation, and posing unacceptable environmental harms.

The recently published EU Code of Practice (EC 2024) details considerations for general-purpose AI models with systemic risks and characterizes “large-scale discrimination” (Section 6.1) under the same category as intolerable risks in Section 2.1 of this paper.

Typically these impacts can only be assessed by monitoring and auditing efforts after models have already been deployed. If unacceptable impacts are discovered from other similar AI models previously deployed, developers should factor unacceptable impacts into their assessment of whether a model breaches their intolerable risk threshold.

2.4 Key Considerations

Determining risk thresholds for dual-use foundation models is an exercise in balancing overarching thresholds with wide applicability versus domain-specific thresholds with high specificity. This specificity may reflect the range of actors, norms, practices, and technical systems involved, and the specific ways in which risks may emerge in that domain (Shelby et al. 2023). Several aspects of frontier model development, model deployment, and other considerations need to be weighed in developing these thresholds (NTIA 2024).

Given the devastating potential of intolerable AI risks, it is imperative to implement policies to prevent them from ever occurring (Ex-Ante) rather than merely implementing safeguards in response to their occurrence (Ex-Post). When developing risk thresholds in this context, empirical research is a highly scarce resource, and it is important to strive for **“good, not**

perfect” thresholds and to err on the side of safety in the face of uncertainty and limited available data.

The following are key considerations to inform the discussion of intolerable risk thresholds:

1. **Additional Risk Criteria:** This working paper has focused on how intolerable risk thresholds are informed by capabilities (in particular), compute, and risks, with additional discussion of how uses, impacts, and limitations may play into intolerable risk thresholds. However, there are other risk criteria that may inform intolerable risk thresholds. For example, in the EU AI Act, additional criteria that inform the designation of general-purpose AI models with *systemic risk* (beyond compute and capabilities) include:
    a. The number of parameters of the model
    b. The quality or size of the data set, for example measured through tokens
    c. The input and output modalities of the model
    d. The size of its reach (e.g. if it will be made available to at least 10,000 business users)
    e. The number of registered end-users
2. **Risk-Benefit Tradeoffs:** Companies may want to compare potentially substantial risks of frontier AI models against potentially substantial benefits. However unacceptable risk is likely to be absolute. In some cases, tradeoffs will be related to relative gains to offensive and defensive capabilities. The assumptions underlying decisions to continue the development of dual-use capabilities need to be explicit. For instance, developing AI capabilities to defend against cybersecurity threats is more promising than developing biological capabilities that are more likely to have a longer timeline to provide satisfactory defensive uses and be at risk of malicious use in the short term. Tentative predictions of offense-defense balance skewing towards offense in increasingly complex AI systems must also be confronted (Shevlane and Dafoe 2020).
3. **Defining Baselines:** To approach sound empirical analyses of model capabilities against thresholds, it is necessary to determine baselines of human performance, other state-of-the-art models, and human-AI systems (UK AISI 2024). Inspired by the threat modeling approaches from computer security, Kapoor, Bommasani et al. (2024) propose a framework to determine empirically sound model evaluation by introducing a framework to assess the **marginal risk** introduced by foundational models when measured against the **baseline of existing threats and defenses** for a particular type of risk. See our Key Principle, "Compare to Appropriate Base Cases" for more information about some of the challenges and limitations of assessing marginal risk.
4. **Deployment Strategy:** Hosted access or restricted access models are often easier to monitor and prevent misuse, whereas the release of model weights presents advantages in terms of access and customization. It is important to assess how different deployment strategies influence the scale, scope, and irreversibility of risks.
5. **Timeframe of Impact:** Current industry discourse on intolerable risks advantages risks that arise from acute catastrophic events that might be made successful through foundation models, like the creation of CBRN weapons or sophisticated cyber attacks

that cause the immediate and massive destruction of life and property. The long-term impacts of frontier models that could fundamentally change the fabric of society are then left for state actors to govern (Lohn and Jackson 2022).

6. **Metrics of Evaluation:** Potential impacts of frontier models can be used to determine the intolerability of risks if there is consensus on the metric of evaluation. For instance, instead of determining intolerance based on the ***'number of human lives lost'***, if we chose to determine impact by measuring its impact on the '***quality of life,'*** our appetites for risk may differ significantly. As an example, a company could set a threshold for carbon emissions from data centers to remain below a certain x% of global/national emissions to prevent climate-related deaths or could apply dynamic metrics to determine the receptiveness of the local population to its continued operations.
7. **Developer vs State Responsibilities:** This working paper focuses on determining a small number of capability thresholds, but we also recommend the consideration of additional intolerable risks (Section 2.3). It is necessary for industry actors to also bear responsibility for these long-term impacts accumulating from frontier model deployment in high-impact systems that operate at scale in critical domains automating tasks, amplifying biases, and increasing emissions which result in long-term effects such as displacing the workforce, amplifying existing inequalities, cultural homogeneity, environmental degradation, etc.
8. **Future-Proofing Benchmarks:**
    a. Intolerable risks do not necessarily require large-scale runs. Therefore, due consideration needs to be placed on how thresholds might have to change rapidly with the widespread availability and affordability of compute power for fine-tuning open models (Seger et al. 2023).

    b. Entrench these margins of safety in threshold determination along the dimensions of increasing affordability, access, and expertise in AI systems to ensure sufficient safety between calibrations.
9. **Benchmarking Often** (or at least at every stage of model development): While model performance can be evaluated against a threshold before its deployment, it is necessary to mimic such robustness in safety testing at every point in the AI pipeline, from data sourcing to determining training sets, to choosing the right hyperparameters, etc. Only by identifying appropriate metrics at each stage of development and evaluating against the right benchmarks can developers exercise responsibility in the product life cycle. Ensuring traceability of decisions and inputs at every stage of development is imperative in enforcing appropriate oversight.
10. **Transparent Evaluations:** These documented risks and decisions should also be reported transparently, to regulators or internal review boards, red teamers, and auditors to ensure appropriate testing against vulnerabilities in the chosen design of the model.

3. Limitations

Identifying thresholds for intolerable risks is one component in enabling AI safety. However, as the Frontier AI Safety Commitments stipulate, there is a simultaneous need for organizations to assume concrete accountability in developing and governing AI systems transparently, as well as allocating sufficient resources to catalyze the development of robust technical tools and techniques to measure and mitigate risks.

1. **Mitigations:** We consider all discussions of capabilities and unacceptable uses in this proposal to apply to the AI models before or without safety guardrails and mitigation measures. This is to simplify the discussion and as an observation that most safety mitigations in use by AI developers today are inadequate or unreliable and can be trivially circumvented via jailbreaks (El-Mhamdi et al. 2022, Wu et al. 2024) or reversed via fine-tuning (Carlini et al. 2023), at least at this point in time (Zou, Wang et al. 2023).
2. **Timeframe in Scope:** This working paper defines certain risk categories and presents additional uses, impacts, and limitations that are intolerable. However, the highest probability of insidious risks from AI deployments will likely be the cumulative effect of ‘high-impact’ AI systems that operate in critical domains automating tasks, amplifying biases, and increasing carbon emissions which result in long-term effects such as displacing the workforce, amplifying existing inequalities, cultural homogeneity, environmental degradation, etc. These risks will require commitments from all actors in the ecosystem, keen monitoring and oversight, and resistance to an inherently tech-first approach to designing access to critical infrastructure and services.
3. **Develop Shared Taxonomies:** This paper builds on frontier model safety frameworks and prominent policy language in categorizing intolerable risks. However, there are a range of taxonomies that need to be standardized to enable uniformity in the risk assessment and reporting exercise. This is where national policies and international agreements can lend their power to operationalize better oversight.
4. **Identify the Right Evaluations**: In accepting current evaluation practices while determining thresholds, there is an implicit assumption of the general reliability of benchmarks. However, model benchmarking and evaluation practices are still nascent and sometimes provide insufficient or incorrect estimations of model capabilities. This should be improved by:
    a. **Construct Validity:** Ground model evaluations in use cases and design evaluation benchmarks curated to the domain, designed with input from the community. (For examples in practice in legal and medical fields, see Guha et al 2023, Nayak et al. 2023)
    b. **Involve End-users:** Citing prompt sensitivity as a crucial challenge, Kapoor, Henderson et al. (2024) recommend the involvement of end users in the evaluation process to adequately determine model capabilities.

References

Anthropic (2024) Responsible Scaling Policy. Anthropic, https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf

Anthropic (n.d.) Responsible Scaling Policy Evaluations Report – Claude 3 Opus. Anthropic, https://cdn.sanity.io/files/4zrzovbb/website/210523b8e11b09c704c5e185fd362fe9e648d457.pdf

Hui Bai, Jan G. Voelkel, Johannes C. Eichstaedt, Robb Willer (2023) Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints*, https://doi.org/10.31219/osf.io/stakv

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, Jessica Newman (2024) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *UC Berkeley Center for Long-Term Cybersecurity*, https://arxiv.org/abs/2405.10986

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, Pascal Vincent (2023a) WorldSense: A Synthetic Benchmark for Grounded Reasoning in Large Language Models. *arXiv*, https://arxiv.org/abs/2311.15930

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, Pascal Vincent (2023b) WorldSense. GitHub, https://github.com/facebookresearch/worldsense

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes (2023a) Emergent autonomous scientific research capabilities of large language models. arXiv, https://arxiv.org/abs/2304.05332

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes (2023b) Autonomous chemical research with large language models. Nature, https://www.nature.com/articles/s41586-023-06792-0

Samantha Bradshaw, Hannah Bailey, and Philip N. Howard (2021) Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation. Oxford Internet Institute, https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/

C4AI (2024) The Limits of Thresholds: Exploring the Role of Compute-Based Thresholds for Governing the Risks of AI Models. Cohere For AI, https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace (2023) Extracting Training Data from Diffusion Models. *arXiv*, https://arxiv.org/abs/2301.13188

Canyu Chen and Kai Shu (2024) Can LLM-Generated Misinformation Be Detected? arXiv, https://arxiv.org/abs/2309.13788

CPR (2023) OPWNAI: Cybercriminals Starting to Use ChatGPT. Check Point Research, https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/

DHS (2011) Regulatory Assessment: The Ammonium Nitrate Security Program Notice of Proposed Rulemaking. US Department of Homeland Security, https://downloads.regulations.gov/DHS-2008-0076-0047/content.pdf

Anca Dragan, Helen King and Allan Dafoe (2024) Frontier Safety Framework. Google DeepMind, https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/

DSIT (2023) Emerging processes for frontier AI. UK Department for Science, Innovation & Technology, https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety

DSIT (2024a) Frontier AI Safety Commitments. AI Seoul Summit 2024. UK Department for Science, Innovation & Technology, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

DSIT (2024b) Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity: AI Seoul Summit 2024. UK Department for Science, Innovation & Technology, https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan

Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone et al. (2024) The Llama 3 Herd of Models. *arXiv*, https://arxiv.org/abs/2407.21783

EC (2024) First Draft of the General-Purpose AI Code of Practice published, written by independent experts. European Commission, https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, Sébastien Rouault, and John Stephan (2022) On the Impossible Safety of Large AI Models. *arXiv*, https://doi.org/10.48550/ARXIV.2209.15259

EP (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). European Parliament, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Max Fisher (2021) Disinformation for Hire, a Shadow Industry, Is Quietly Booming. New York Times, https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html

Lukas Fluri, Daniel Paleka, and Florian Tramèr (2023). Evaluating superhuman models with consistency checks. arXiv: https://arxiv.org/abs/2306.09983

FMF (2024) Issue Brief: Early Best Practices for Frontier AI Safety Evaluations. Frontier Model Forum, https://www.frontiermodelforum.org/updates/early-best-practices-for-frontier-ai-safety-evaluations/

Victoria A. Greenfield, Henry H. Willis, Tom LaTourrette (2012) Assessing the Benefits of U.S. Customs and Border Protection Regulatory Actions to Reduce Terrorism Risks. RAND Corporation,
https://www.rand.org/content/dam/rand/pubs/conf_proceedings/2012/RAND_CF301.pdf

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova (2023) Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv, https://arxiv.org/abs/2301.04246

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li (2023) LegalBench: A

Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models *arXiv*, http://arxiv.org/abs/2308.11462

Lennart Heim and Leonie Koessler (2024) Training Compute Thresholds: Features and Functions in AI Regulation. arXiv, https://arxiv.org/pdf/2405.10799

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2020) Measuring Massive Multitask Language Understanding. *arXiv*, https://arxiv.org/abs/2009.03300

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, Ethan Perez (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv, https://arxiv.org/abs/2401.05566

Olli Järviniemi and Evan Hubinger (2024). Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant. arXiv: https://arxiv.org/abs/2405.01576

JCS (2018) Operations in Chemical, Biological, Radiological, and Nuclear Environments. Joint Chiefs of Staff, https://www.jcs.mil/portals/36/documents/doctrine/pubs/jp3_11.pdf

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins et al. (2024) On the Societal Impact of Open Foundation Models. *arXiv*, https://arxiv.org/abs/2403.07918

Sayash Kapoor, Peter Henderson, and Arvind Narayanan (2024) Promises and Pitfalls of Artificial Intelligence for Legal Applications. *arXiv*, https://arxiv.org/abs/2402.01656

Sayash Kapoor, Arvind Narayanan (2023) How to Prepare for the Deluge of Generative AI on Social Media. Knight First Amendment Institute, https://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media

Holden Karnofsky (2024) A Sketch of Potential Tripwire Capabilities for AI [Unpublished draft]. Carnegie Endowment for International Peace.

Leonie Koessler, Jonas Schuett, and Markus Anderljung (2024) Risk thresholds for frontier AI. arXiv, https://arxiv.org/abs/2406.14713

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, Owain Evans (2024). Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. arXiv, https://arxiv.org/abs/2407.04694

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024a) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *arXiv*, https:/arxiv.org/abs/2403.03218

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024b) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *GitHub*, https://github.com/centerforaisafety/wmdp

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024c) Datasets: cais/wmpd. *Hugging Face*, https://huggingface.co/datasets/cais/wmdp

Zilong Lin, Jian Cui, Xiaojing Liao, XiaoFeng Wang. (2024). Malla: Demystifying Real-world Large Language Model Integrated Malicious Services. arXiv, https://arxiv.org/abs/2401.03315

Andrew Lohn and Krystal Jackson (2022) Will AI Make Cyber Swords or Shields? Center for Security and Emerging Technology, https://perma.cc/3KTH-GQTG

Sandra Matz, Jake Teeny, Sumer S. Vaid, Heinrich Peters, Gabriella M. Harari, and Moran Cerf (2024) The Potential of Generative AI for Personalized Persuasion at Scale. Nature, https://www.nature.com/articles/s41598-024-53755-0

Christopher A. Mouton, Caleb Lucas, and Ella Guest (2024) The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. *RAND Corporation*, https://www.rand.org/pubs/research_reports/RRA2977-2.html

Micah Musser (2023) A Cost Analysis of Generative Language Models and Influence Operations. arXiv, https://arxiv.org/abs/2308.03740

Ashwin Nayak, Matthew S. Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P. Weinfurt, and Kevin Schulman (2023) Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents. *JAMA Network*, https://doi.org/10.1001/jamainternmed.2023.2561.

NTIA (2024) Dual-Use Foundation Models with Widely Available Model Weights. National Telecommunications and Information Administration, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

OMB (2024) Memorandum for the Heads of Executive Departments and Agencies. Office of Management and Budget, https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf

OpenAI (2023a) GPT-4 System Card. OpenAI, https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI (2023b) Preparedness Framework (Beta). OpenAI, https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenAI (2024) Disrupting deceptive uses of AI by covert influence operations. OpenAI, https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, Dan Hendrycks (2023). AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv, https://arxiv.org/abs/2308.14752

Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn (Froggi) Jackson, Steven Adler, Rocco Casagrande, and Aleksander Madry (2024) Building an early warning system for LLM-aided biological threat creation. *OpenAI*, https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation

Stephen R. Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, Liam G. McCoy, Leo Anthony Celi, Yun Liu, Mike Schaekermann, Alanna Walton, Alicia Parrish, Chirag Nagpal, Preeti Singh, Akeiylah Dewitt, Philip Mansfield, Sushant Prakash, Katherine Heller, Alan Karthikesalingam, Christopher Semturs, Joelle Barral, Greg Corrado, Yossi Matias, Jamila Smith-Loud, Ivor Horn, and Karan Singhal (2024) A toolbox for surfacing health equity harms and biases in large language models. arXiv, https://arxiv.org/abs/2403.12025

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Grégoire Delétang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane (2024) Evaluating Frontier Models for Dangerous Capabilities. *arXiv*, https://arxiv.org/abs/2403.13793

Andrew Rae, Rob Alexander, John McDermid (2014) Fixing the cracks in the crystal ball: A maturity model for quantitative risk assessment, Reliability Engineering & System Safety, https://doi.org/10.1016/j.ress.2013.09.008

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider (2024) Escalation Risks from Language Models in Military and Diplomatic Decision-Making. arXiv, https://arxiv.org/pdf/2401.03408

Stuart Russell (2024) Make AI safe or make safe AI? UC Berkeley, https://aima.cs.berkeley.edu/~russell/papers/russell-unesco24-redlines.pdf

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West (2024) On the conversational persuasiveness of large language models: A randomized controlled trial. arXiv, https://arxiv.org/abs/2403.14380

Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan (2021) AI and the Future of Disinformation Campaigns, Part 2: A Threat Model. Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta (2023) Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. arXiv, https://arxiv.org/abs/2311.09227

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk (2023)

Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv, https://arxiv.org/abs/2210.05791

Toby Shevlane and Allan Dafoe (2020) The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? arXiv, https://arxiv.org/abs/2001.00463

Almog Simchon, Matthew Edwards, Stephan Lewandowsky (2024) The persuasive effects of political microtargeting in the age of generative artificial intelligence, https://doi.org/10.1093/pnasnexus/pgae035

Cass R. Sunstein (2006). Irreversible and Catastrophic. Cornell Law Review, http://scholarship.law.cornell.edu/clr/vol91/iss4/2

UK AISI (2024) Early Lessons from Evaluating Frontier AI Systems. UK AI Safety Institute, https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022a) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *arXiv,* https://arxiv.org/abs/2206.10498

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022b) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *GitHub*, https://github.com/karthikv792/LLMs-Planning/tree/main/plan-bench

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, Francis Rhys Ward (2024). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. arXiv, https://arxiv.org/abs/2406.07358

White House (2023) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Executive Order 14110. White House, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

White House (2024) Framework to Advance AI Governance and Risk Management in National Security. White House, https://ai.gov/wp-content/uploads/2024/10/NSM-Framework-to-Advance-AI-Governance-and-Risk-Management-in-National-Security.pdf

Angus R. Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E. Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright (2024) Large language models can consistently generate high-quality content for election disinformation operations. arXiv, https://www.arxiv.org/abs/2408.06731

Baoyuan Wu, Zihao Zhu, Li Liu, Qingshan Liu, Zhaofeng He, and Siwei Lyu (2024) Attacks in Adversarial Machine Learning: A Systematic Survey from the Life-cycle Perspective. *arXiv*, https://arxiv.org/abs/2302.09457

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*, https://arxiv.org/abs/2307.15043

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan et al. (2023). Representation engineering: Understanding and controlling the inner workings of neural networks. arXiv:https://arxiv.org/abs/2310.01405

Appendices

Appendix A: Types of Thresholds for Frontier Models

Risk Thresholds

Risk is often defined in terms of likelihood (probability of an event), and severity of harm (magnitude of impact). Comprehensive risk models containing all possible risk scenarios are extremely difficult to develop, and it is recommended to start with a limited and defined number of risk scenarios (Koessler et al. 2024). Regardless of the risk measurement method, it is important that risk thresholds are operationalized and are specific enough such that multiple evaluators with access to the same resources would agree on the risk threshold determination of an evaluated model (DSIT 2023). A common proxy used by model providers to measure risk is model capability (see Anthropic 2024, Dragan et al. 2024). Another common proxy measure for risk is compute power, defined in both Executive Order 14110 and the EU AI Act.

Compute Thresholds

Compute thresholds are often measured with floating-point operations per second (FLOPS). The White House Executive Order 14110 (White House 2023) and the EU AI Act (EP 2024) have both already made use of compute thresholds to categorize high-risk models. Compute thresholds may be most useful as an initial metric to identify models that require further regulatory oversight and evaluation. Compute thresholds set above the frontier (currently 10^{26} operations) would include models that have dangerous capabilities that are difficult to predict. Compute thresholds set at the frontier (currently 10^{25} operations) would include models that may already show warning signs of dangerous capabilities. Compute thresholds set below the frontier (currently 10^{24} operations) would be the most cautious approach, creating a larger safety margin, but may create unnecessary regulatory burdens (Heim and Koessler 2024). In addition to compute thresholds, it is recommended to utilize alternative approaches and develop dynamic thresholds. There is also a need to clearly specify approaches for measuring FLOPs for various types of systems (C4AI 2024).

Capability Thresholds

Model capability can serve as an (imperfect) proxy for risk and is considered a key determinant of risk. Because of the dual-use nature of foundation models, risk often stems from capability (Koessler et al. 2024). These advanced models may amplify societal risks if they are exploited to increase the effective ability of malicious actors to execute attacks, or are deployed to autonomously execute attacks (e.g. cyber and CBRN attacks) (Barrett et al. 2024, UK AISI 2024). At least two methods are available to evaluate model capability: open benchmarks, and closed red teams. Benchmarks utilize a standardized set of questions and answers through model prompts to evaluate model capability, making them a quick and cost-effective option. An example of a CBRN and cyber-related benchmark is WMDP (Li et al. 2024a,b,c). Other capability benchmarks include PlanBench (Valmeekam 2022a,b) WorldSense (Benchekroun et al. 2023a,b), and MMLU (Hendrycks et al. 2020). While quick and cost-effective, benchmarks lack accuracy. Red team evaluations involve intensive and interactive testing by domain experts and achieve a higher level of accuracy by incorporating sensitive details. For more on red teams see, RAND (Mouton et al. 2024) and OpenAI (Patwardhan et al. 2024).

Appendix B: Examples of Quantitative Capability Thresholds

Here we provide examples of relatively specific, numerical thresholds for model capabilities that we expect could inform reasonable intolerable risk thresholds. We note that numerical thresholds hold potential benefits for accountability, while also acknowledging the limitations of quantitative risk assessment (Rae et al. 2014).

We recommend considering these thresholds as part of a starting point for creating a broader and more in-depth set of intolerable thresholds. Readers also should keep in mind that we regard these as examples of possible reasonable thresholds; other, similar thresholds may also be reasonable. In addition, it may be appropriate to use these as initial thresholds, and to revise them later (e.g., in response to new information on offense-defense balance in a particular domain).

Case 1: CBRN Weapons: Example Capability Thresholds

In the following CBRN capability threshold example, we aim to operationalize intolerable risk thresholds at approximately the “substantial” level. That would be higher than a de-minimis “detectable” level. However, it also could be somewhat lower than at a “severe” level; it would be prudent to leave some margin of safety before arriving at a “severe” level.[7] We also provide a few comments on the general reasoning we used in creating these thresholds, and on how they compare to some CBRN model capability evaluation methods and/or thresholds used by industry.

A foundation model should be regarded as presenting a substantial lowering of barriers to CBRN attack, and an intolerable risk, if, for any relevant CBRN attack stage, the model provides any of the following, where applicable:

[7] For comparison, our provisional recommendations are to regard as substantial lowering of barriers an effect approximately equivalent to Anthropic’s Yellow Line for CBRN – an 25% absolute increase in accuracy – and to regard as severe lowering of barriers an effect approximately equivalent to Anthropic’s Red Line for CBRN – allowing a non-expert to reach human expert levels (Anthropic n.d.). Note that Anthropic described the Red Line as “substantial lowering of barriers”, implying a higher risk tolerance than we recommend.

- For models with CBRN capability effects rated in terms of **probability of adversary success in carrying out a CBRN attack**, if the adversary attempts that attack:[8]
    - The model provides **an absolute increase in adversary success probability of at least 25%**
        - For increasing estimated adversary probability of success given attempt (where that probability is estimated on a 0% to 100% scale), if that is part of the rating process
- For models with CBRN capability effects rated in terms of attack plan **accuracy**, **completeness,** or other key technical or operational dimensions:**[9]**
    - The model provides **an absolute increase in those dimensions of at least 25%**
        - For lowering of barriers towards an operational plan with sufficient accuracy, completeness, etc. to be viable (where such technical or operational dimensions are rated using a constructed scale ranging from 0% to 100%, or analogous scales that can be normalized to approximate equivalents to a 0-100% scale)
- For models with CBRN capability effects rated **in other terms, in comparison to human expert levels:**
    - The model **lowers barriers by half (i.e. getting halfway from baseline non-expert to human expert levels in evaluations)[10]** if that represents at least 10% absolute increase (i.e. a greater than minimal effect)

[8] Pre-release red teaming of Llama 3 included evaluation of the model's chemical and biological capabilities in terms of human uplift effect on **adversary probability of success**. "....Participants were asked to generate fictitious operational plans for either a biological or chemical attack…. Each team is assigned to a "control" or "LLM" condition. The control team has access to internet-based resources only, while the LLM-enabled team had internet access as well as access to Llama 3 models …. At the conclusion of the exercise, the operational plans generated by each team are evaluated by subject matter experts with domain expertise in biology, chemistry, and operational planning. Each plan is evaluated across four stages of potential attacks, generating scores for metrics such as scientific accuracy, detail, detection avoidance, and probability of success in scientific and operational execution." (Dubey et al. 2024, p. 47)

[9] Examples of frameworks for, or implementation of, pre-release red teaming of models that included evaluation of a model's biological capabilities in terms of human uplift effect on **accuracy** of biological attack operational plans include: the pre-release red teaming of Llama 3 (Dubey et al. 2024, p. 47; the pre-release red teaming of Claude (Anthropic n.d.).

[10] This would be an example of operationalizing intolerable risk thresholds at approximately the "substantial" level for lowering of barriers to CBRN attack, where a model's human-uplift effect is half of the difference between human non-expert baseline and human expert. That leaves some margin of safety before arriving at a human-expert or "severe" level for lowering of barriers. Leaving such a margin of safety would be prudent, especially for models intended for open weights release that would be easiest to fine tune, or to enhance in other ways such as reinforcement learning and chain of thought to add capabilities, but that cannot be monitored or decommissioned by the model developer through an API.

Case 2: Deception: Example Capability Threshold

As discussed earlier, although deception ultimately poses intolerable risks we want to detect and prevent, we expect it will be extremely difficult to evaluate deception directly in an AI model based on a deep neural network architecture. Hence precursor capabilities for deception such as situational awareness become relevant. (Laine et al. 2024, pp. 33-34)

While data is particularly limited for this risk category, we propose the following intolerable risk threshold for deception:

Sufficient situational awareness for the model to correctly identify whether its current environment is evaluation or deployment with 60% or greater probability.

Rationale: A model with no situational awareness of its runtime environment should correctly distinguish between evaluation and deployment environments 50% of the time. A model reaching 60% awareness of the environment has a non-negligible ability to strategically underperform (sandbag) on capability evaluations, *including situational awareness evaluations* (van der Weij et al. 2024). Hence situational awareness needs to be evaluated frequently during gradual model training in order to detect possibilities of deception, and if a significant level (we propose 60%) of situational awareness is reached during training, evaluations may be less trustworthy and models will require additional scrutiny.

Appendix C: Key Principles to Determine Thresholds

1. Seek to Identify Cases of Substantial Increase in Risk

We use the principle of giving higher priority to risks with a relatively high magnitude of effect, especially where impacts could be regarded as irreversible or catastrophic (Sunstein 2006).

We also briefly introduce the basic concept of marginal risk in Section 2.4, as well as the basic principle of risk management to prioritize risks based on the size of marginal risk. At a minimum, we aim to identify cases of substantial increase to marginal risk. The concept of marginal risk can be a useful way to compare the risks of a model compared to standard tools such as typical searches on the internet, but the concept can also contribute to a slippery slope environment where progressively worse and more dangerous models are considered acceptable. It is not appropriate to compare the risks of a new frontier model to the risks of every other available model because there are already widely proliferated models without sufficient safeguards that can be used to cause significant harm. (We return to that point later, in this subsection in the “Compare to Appropriate Base Cases” principle).

We assume several terms are approximately equivalent to “substantial”. We do not perform a legal analysis of such terms here. We assume that “substantial” can mean something greater than “detectable”.

2. Focus on Capabilities, at Least Until Likelihood Estimation Becomes More Reliable

Koessler et al. (2024) have argued for defining intolerable risk thresholds primarily in terms of “risk”, i.e., which include both likelihood and consequence, then using such intolerable risk thresholds to inform capability thresholds. In principle, it seems very logical to use that approach, and that methods such as benefit-cost analysis should inform risk thresholds.

However, the literature on risk management for rare or novel catastrophic events, such as terrorism, suggests that applying cost-benefit analysis to such domains often introduces debatable assumptions or at least must factor in great uncertainties (see, e.g. discussion of break-even analysis in DHS 2011 and Greenfield et al. 2012). That does not so much indicate where to draw a clear threshold "line"; instead it suggests a very broad plausible range.

Thus, at the current point in time and for the next few years, it seems substantially more reliable to use capability thresholds (more so than compute thresholds) without compounding uncertainties of likelihood estimates, while still taking impact into account.

3. Compare to Appropriate Base Cases

Typically with human-uplift studies, CBRN and cyber dual-use capability assessment methods either implicitly or explicitly compare a model’s outputs to information available from Web searches. (See, e.g., Mouton et al. 2024, Patwardhan et al. 2024, and Anthropic n.d. on bio domain comparisons, and Dubey et al. 2024 on cyber as well as CBRNE.)[11] This seems appropriate, especially when considering the risks of models lowering barriers to CBRN or cyber weapons for the relatively large numbers of potential low-technical capability adversaries that lack high baseline technical education or other technical capabilities in a particular domain.

In addition to comparing a model’s outputs to information from the Web, it also could be useful and appropriate to compare a model’s outputs to information from domain-specific textbooks or other technical documents that are not available on the open Internet or to evaluate a model’s capabilities for lowering barriers to the use of biological design tools and/or lab automation functions. That would be useful when considering the risks of models lowering barriers to especially high-consequence CBRN or cyber weapons, such as novel or enhanced pandemic potential pathogens, for high-technical capability adversaries.[12]

For assessing the marginal risks of releasing a particular model, it could be valuable to compare a new LLM to other available LLMs, instead of to the Web. However, a model’s outputs should

[11] The October 2024 Responsible Scaling Policy update (Anthropic 2024) specifically mentions information on the Web circa 2023.

[12] For more considerations for CBRN and cyber threat modeling, see discussion and references in Section 2 of Barrett et al. (2024).

not only be compared to other available models. Closed-weights models can be rolled back, and they can be made unavailable very quickly via the provider's control of an API, but open-weights models cannot effectively be made unavailable after the release of their weights. With growing investment in AI globally and an increasing number of models released each year, using existing models as a baseline could easily lead to an exponential growth in risk from AI models overall, year over year. Moreover, open-weights models should not be provided with an easy path to releasing ever more powerful frontier models via an incrementally rising marginal-risk comparison to their last open-weights release. That would be a slippery slope and bad risk management public policy.

4. Minimal Increases to Risk Should be Detectable, but not Necessarily Intolerable

AI developers and evaluators should not be disincentivized for good-faith measurement efforts that detect a very small level of capabilities. Indeed, there is substantial value in constructing evaluation processes that are sensitive enough to detect small levels of capabilities.

Thus, intolerable risk thresholds should not be so low as to imply that intolerable risks include "anything detectable by any means available", or "anything statistically significant". It's possible to have statistically significant effects that have a small magnitude of effect.

An example of something that could be "anything detectable" that could be less than "significant" would be a 10% increase on a single dimension like accuracy, completeness, or other key technical or operational dimensions. At least for CBRN, it seems unrealistic to assess accuracy on anything more granular than the nearest 10%. (Cyber benchmarks may provide greater granularity, but it seems reasonable to use this same general principle for cyber as for CBRN.)

Note that small capabilities should not be merely ignored, instead, they should be used as potential indicators for other hazardous capabilities, and as triggers for additional efforts to detect risk more broadly or in more depth. This is typically what responsible scaling policies and similar policies do, and should aim to do.

5. A Substantial Increase in Capabilities for Some Attack Stages Can Have a Disproportionately Large Effect on Risk

For many CBRN attack scenarios, one or two stages are the main limiting factor affecting an adversary's chance of success. For example, for nuclear weapons, obtaining sufficient quantities of fissile material is typically the main limiting factor. If an adversary's baseline chance of success in each of several attack stages is 75% in those attack stages, then a 25% absolute increase in chance of success in any one of those stages would represent a 33% increase in their overall chance of success – likely to seem helpful, but not game-changing, to many adversaries. By contrast, if an adversary's baseline chance of success for a particular attack scenario is 25% in a particular stage, then a 25% absolute increase in that one stage would represent a 100% increase in their overall chance of success. Such increases may seem game changing to many adversaries, even if their resulting chances of success are below 100%.[13] That would not only increase the probability of success if an adversary attempts an attack scenario, but it also likely would increase the number of adversaries that would attempt a scenario, because they would feel more likely to succeed.

6. Leave Some Margin of Safety

Anthropic defined redline CBRN capabilities in terms of getting all the way to human-level expertise. However, it is worth leaving a margin of safety, especially for models intended for open weights release that would be easiest to fine-tune, or to enhance in other ways such as reinforcement learning and chain of thought to add capabilities but that cannot be monitored or decommissioned by the model developer through an API. Thus, it’s worth aiming to stay well below that, e.g. below a threshold of lowering barriers by half (i.e. getting halfway to human expert levels).

More generally, it seems prudent to operationalize intolerable risk thresholds at approximately the “substantial” level, leaving some margin of safety before arriving at a “severe” level.

7. Define at Least Some Thresholds Without Factoring in Model-Capability Mitigations

Some unacceptable model capability thresholds should not factor in model guardrails or other model-capability mitigation measures. Virtually all guardrails for capability are unreliable (e.g. via jailbreaks) (El-Mhamdi et al. 2022, Wu et al. 2024) or reversible (e.g. via fine-tuning) (Carlini et al. 2023), at least at this point in time (Zou, Wang et al. 2023).

intolerable risk thresholds also should reflect the degree to which societal mitigations are feasible. For the initial operationalization of intolerable risk thresholds, it may be appropriate to focus first on CBRN rather than cyber. One reason is that new physical defenses (including against CBRN attacks) are often harder, more expensive, and more time-intensive than software patches (including against cyber-attacks). Thus, cyber offensive capabilities may have more value than CBRN offensive capabilities for defenders seeking to identify vulnerabilities to patch. These could be reasons to define a bright line earlier for CBRN than for cyber where it may make more sense to take a more adaptive approach.

8. What if an Uncertainty Range Overlaps with a Threshold?

Depending on the statistical approach and risk management decision rules used, AI developers may aim to evaluate whether 1) the mean value of uplift is less than a threshold, or 2) the upper end of the confidence interval for the value of uplift is less than a threshold. If a confidence interval overlaps with the intolerable risk threshold, then that should be a reason to do more

[13] A senior terrorism expert privately told one of us years ago that an archetypal terrorist would seriously consider attempting an attack scenario if that adversary thought their chance of success for that scenario was at least 75%. This mental model is obviously an oversimplification, but we see it as close enough to reality to usefully inform our thinking in this context.

work with the larger sample size to reduce the uncertainty range, to give a better assessment of whether the effect is actually below the intolerable risk threshold.

Use Best Practices in Dual-Use Capability Evaluation

Thresholds are only meaningful in the context of a rigorous capability evaluation process. These processes should include reasonable good-faith use of best practices, including:

- Enough relevant scenarios (CBRN agents/materials; threat actor capability levels; etc.) to sufficiently sample the space of key scenarios
- Deploying diverse assessment methodologies (Pfohl et al. 2024)
- Large enough participant sample sizes
- Red teamer access to versions of models that do not require jailbreaking
- Methods to assess a model's capabilities for situational awareness, sandbagging, or other capabilities for deception that could lead to evaluators underestimating a model's CBRN, cyber, or other dual-use capabilities
- Red teamer ability to perform reasonably foreseeable capability enhancements such as plugin tools (especially for cyber) or fine tuning (especially for CBRN), either to remove safety filters or to add capabilities by training on domain specific corpora such as on CBRN
    - This is important for closed release models that will be released with fine-tuning access, and especially important for models intended for open weights release for which fine-tuning will be especially easy
    - Cyber capabilities work can and should include plug-in tools and scaffolding when evaluating capabilities (see, e.g., Phuong et al. 2024)