WHITE PAPER
August 2025

AI GOVERNANCE INITIATIVE

OXFORD MARTIN SCHOOL

UNIVERSITY OF OXFORD

SURVEY ON THRESHOLDS FOR ADVANCED AI SYSTEMS

Authors: Jonas Schuett, Eunseo Choi, Kasumi Sugimoto, Bosco Hung, Robert Trager, and Karine Perset

Centre for the Governance of AI

Survey on thresholds for advanced AI systems

Jonas Schuett[*1] **Eunseo Choi**[2] **Kasumi Sugimoto**[2]

Bosco Hung[3] **Robert Trager**[4] **Karine Perset**[2]

[1]Centre for the Governance of AI [2]OECD.AI
[3]University of Oxford [4]Oxford Martin AI Governance Initiative

Abstract

Governments around the world have recognised the need to manage risks from advanced artificial intelligence (AI) systems. Thresholds are discussed as a potential governance tool that could be used to determine when additional risk assessment or mitigation measures are warranted. However, it remains unclear what specific thresholds would be appropriate and how they should be set. This paper reports findings from an expert survey (N = 166) and a public consultation conducted between August 2024 and October 2024. The expert survey asked participants to indicate their level of agreement with 98 statements about thresholds for advanced AI systems. The public consultation provided an opportunity for the general public to contribute perspectives that may not have been captured by the expert survey. Participants generally agreed that thresholds should be set by multiple stakeholders and that there should be different types of threshold, each serving a specific purpose. Participants were divided on the question of what role training compute thresholds should play, how exactly different types of thresholds should be set, and how many thresholds there should be. These findings can serve as evidence in ongoing policy discussions, yet more research is needed.

[*]Corresponding author: jonas.schuett@governance.ai.

Contents

Executive Summary

A number of documents published at the AI Seoul Summit 2024, including the Ministerial Statement, the Frontier AI Safety Commitments, and the International Scientific Report on the Safety of Advanced AI, indicate a growing interest in thresholds for advanced AI systems. Thresholds also play a key role in AI regulations (e.g. the EU AI Act and GPAI Code of Practice) and in AI risk management standards (e.g. the NIST AI Risk Management Framework and ISO/IEC 23894). However, it remains unclear what specific thresholds would be appropriate and how they should be set.

This paper reports findings from an expert survey ($N = 166$) and a public consultation conducted between August 2024 and October 2024. The expert survey asked participants to indicate their level of agreement with 98 statements about thresholds for advanced AI systems. The public consultation provided an opportunity for the general public to contribute perspectives that may not have been captured by the expert survey.

The expert survey and public consultation identified several areas of agreement and divergence among experts. Participants generally agreed that:

- **Thresholds should be set by multiple stakeholders.** They should not be set solely by AI companies or governments.
- **There should be multiple types of threshold, each serving a specific purpose.** For example, different types of thresholds may be used to determine whether an AI system requires further scrutiny, whether additional mitigations are warranted, whether the development process should be paused, or whether the system can be deployed. Different types of thresholds should be based on different metrics, such as model capabilities, training compute, estimates of the probability and severity of harm, deployment context, or an AI system's level of autonomy. Importantly, not all thresholds should necessarily be model-based.
- **Thresholds should have certain qualities.** They should be verifiable by external actors and enforceable by government authorities. They should also be future-proof, justified, and account for uncertainties.
- **The process of setting and evaluating thresholds involves many challenges.** In particular, the process can be influenced by conflicts of interest. AI companies may underestimate risks, whether intentionally or unintentionally, to stay below risk thresholds. Measuring metrics for these thresholds also presents challenges. Reliable methods for estimating risks are lacking, making it difficult to determine when risk thresholds are exceeded. Similarly, setting and evaluating capability thresholds – focused on model capabilities and adequate mitigations – is complicated by current limitations in model evaluations.
- **AI companies should report when thresholds are exceeded.** Most participants agreed that, if any thresholds are breached, AI companies should disclose the breach publicly and inform key stakeholders. Besides that, experts did not agree on what measures AI companies should take when thresholds are breached.

Participants were divided on the following questions:

- **What role should training compute thresholds play?** Training compute thresholds are thresholds defined in terms of the computational resources used to train a model. While some participants highlighted advantages of training compute thresholds (e.g. they are easy to measure and verifiable by external actors), others raised concerns (e.g. they are gameable and can quickly become ineffective). Despite these diverging views, experts widely agreed that training compute thresholds should not be used alone to determine whether an AI system poses unacceptable risks because training compute is an imperfect proxy for risk.
- **How should different thresholds be set?** Although participants generally agreed that thresholds should be justified, there was no consensus on how specific thresholds should be justified.
- **How many thresholds should there be?** While participants generally agreed that there should be different types of thresholds, they were uncertain about how many tiers of thresholds there should be. Several participants flagged that the optimal number of thresholds depends on various contextual factors.

- **Are thresholds an appropriate governance tool?** Some participants questioned whether thresholds are an appropriate governance tool for advanced AI systems, given the the methodological challenges involved.

Findings from this study represent a snapshot of views at a moment in time. It is important to keep in mind that variations in how key terms were interpreted and the subjective nature of the qualitative analysis, both underscore the need for caution when interpreting the results. Moreover, while the survey was designed and reviewed externally to explore several different claims, it was not intended to capture the full spectrum of views within the AI community and the general public. The survey also did not explicitly ask participants whether or for which problems thresholds are appropriate. As a result, the survey may have implicitly suggested that thresholds are desirable. Similarly, including questions about compute and capability thresholds may have suggested that thresholds should focus on certain model properties.

Despite these limitations, the study reveals areas of agreement and divergence that extend beyond individual viewpoints. These insights are useful for understanding the complexities of setting and evaluating AI thresholds and can inform ongoing policy discussions.

Figure continues on next page

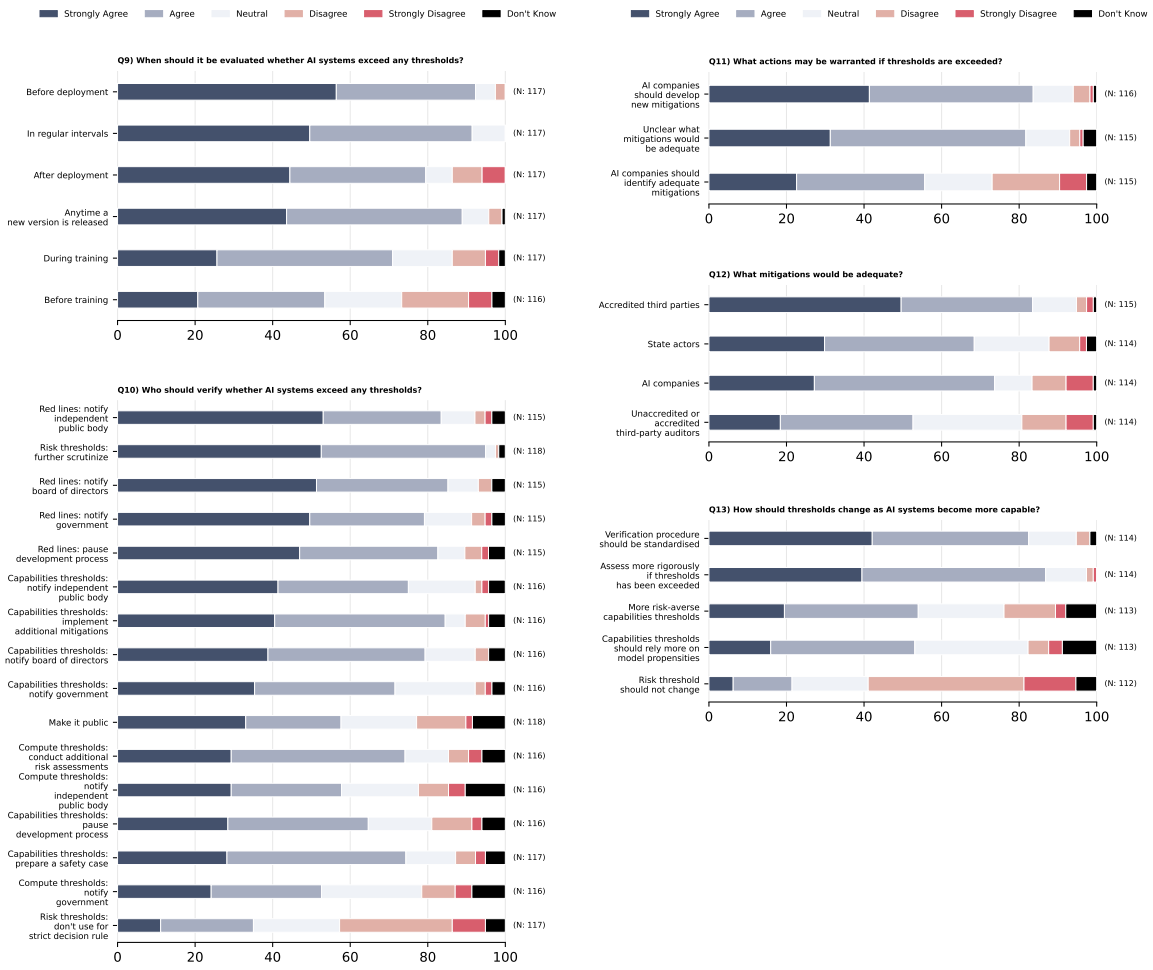

Figure 1: Percentages of responses for all questions and statements

1 Introduction

At the first global AI Safety Summit hosted by the United Kingdom in 2023, 28 governments signed the Bletchley Declaration, which recognises the urgent need to understand and collectively manage potential risks from advanced AI systems (DSIT, 2023). This need is also recognised by the International Guiding Principles for Organisations Developing Advanced AI Systems, adopted at the G7 Hiroshima Summit 2023 (G7, 2023). Building on this recognition, a number of documents published at the AI Seoul Summit 2024, including the Ministerial Statement (DSIT, 2024b), the Frontier AI Safety Commitments (DSIT, 2024a), and the International AI Safety Report (Bengio et al., 2025), indicate a growing interest in thresholds[1] for advanced AI systems.[2] Thresholds are also central in AI regulations such as the EU AI Act (European Parliament, 2024) and the GPAI Code of Practice (European Commission, 2025a) as well as in AI risk management standards like the NIST AI Risk Management Framework (NIST, 2023) and ISO/IEC 23894 (ISO & IEC, 2023). However, it remains unclear what specific thresholds would be appropriate and how they should be set.

While there is extensive literature on risk thresholds in other industries (Linkov et al., 2011; Marhavilas & Koulouriotis, 2021; Fischhoff et al., 1981; Klinke & Renn, 2002; Starr, 1969), research specifically addressing thresholds for advanced AI systems is still emerging (Koessler et al., 2024; Heim & Koessler, 2024; Hooker, 2024; Raman et al., 2025; Caputo et al., 2025). This paper seeks to contribute to this body of literature by identifying areas of agreement and divergence among experts from academia, civil society, as well as the private and public sector. It reports findings from an expert survey and a public consultation conducted between August 2024 and October 2024.

The paper is organised as follows. Section 2 describes the methods used for the expert survey and the public consultation. Section 3 reports some of the main results. Section 4 discusses general results that apply to all thresholds and specific results that only apply to certain types of thresholds. It also flags key limitations of the study and suggests directions for future research. Section 5 concludes. Appendix A lists all survey participants who consented to have their names and affiliations mentioned, while Appendix B lists all consultation participants. Appendix C lists all statements used in the survey.

2 Methods

This section outlines the methods used for the expert survey (Section 2.1), public consultation (Section 2.2), and data triangulation (Section 2.3).

2.1 Method for the expert survey

The survey was conducted between mid-September and mid-October 2024. Invitations and reminders were sent via email, and the deadline was extended three times to accommodate more participants.

2.1.1 Survey questionnaire

Survey design. The survey began with background information on thresholds for advanced AI systems, including definitions of key terms (see Table 1), clarification of the types of risk being considered, instructions for participation, and details on the protection of personal data. Experts were then asked to indicate their level of agreement or disagreement with various statements regarding thresholds for advanced AI systems. Participants also had the opportunity to provide their rationales or raise additional considerations. All questions were optional. The survey was hosted on LimeSurvey.

Statements about thresholds for advanced AI systems. The survey included 98 statements (see Appendix C), organised into 13 questions (see Table 3). Participation in answering these questions was optional. To improve the credibility and clarity of the survey, as well as to check for bias and ensure the relevance of the statements, feedback and suggestions for improvement were sought from researchers at non-governmental research institutions in various geographical locations (Bengio et al.,

[1] Note that there are several related terms such as 'risk acceptance criteria", risk tolerance", risk appetite", and risk tiers". While some scholars think that there are conceptual differences among these terms, others treat them as mostly synonymous.

[2] Sometimes referred to as general-purpose AI", frontier AI", or foundation models".

Term	Definition	Literature
Advanced AI	Highly capable AI models or systems that can perform a wide variety of tasks	G7 (2023), Bengio et al. (2025)
Risk	Combination of the probability of occurrence of harm and the severity of that harm	European Parliament (2024), NIST (2023)
Threshold	Predefined point above which additional mitigations are deemed necessary	Koessler et al. (2024)
Risk threshold	Threshold defined in terms of the probability and severity of harm	Koessler et al. (2024), Raman et al. (2025), Caputo et al. (2025)
Capability threshold	Threshold defined in terms of model capabilities and adequate mitigations*	Koessler et al. (2024), Frontier Model Forum (2025c), METR (2025)
Red lines	Threshold defined in terms of unacceptable model capabilities regardless of mitigations**	Bengio et al. (2024), IDAIS (2024), Zoumpalova and Iliadis (2025)
Training compute threshold	Threshold defined in terms of the computational resources used to train a model	Heim and Koessler (2024), Hooker (2024), Koessler et al. (2024), Sastry et al. (2024); Frontier Model Forum (2024a), Pistillo et al. (2025), Cottier and Owen (2025)

* Note that the definition of the term "capability threshold" is not accepted universally and may be potentially confusing. While some practitioners use the term to refer to capability-mitigation mappings, it could also be interpreted as referring solely to capabilities. Under the latter interpretation, mitigations would be a response to exceeding capability thresholds rather than an inherent defining property of the threshold itself.

** It is important to note that there is no universally accepted definition or conception of the term "red lines".

Table 1: Key terms used in the survey

2025), and two partners drafted policy papers on training compute thresholds and risk thresholds for AI (Heim & Koessler, 2024; Koessler et al., 2024).

Response scale. Participants were asked to indicate their level of agreement based on a 5-point Likert scale (Likert, 1932): 'strongly disagree" (-2), somewhat disagree" (-1), neither agree nor disagree" (0), somewhat agree" (1), strongly agree" (2). They also had the option to say I don't know" (5). This scale is commonly used in expert elicitation studies (B. Zhang & Dafoe, 2020; B. Zhang et al., 2021; Schuett et al., 2023).

2.1.2 Sample

Sample size. Out of 379 experts invited to participate in the survey, 166 experts participated (43.8% participation rate), though not everyone answered all optional questions.

Sample selection. The survey was distributed to members of the OECD.AI expert community. This informal group consists of experts from government, business, academia, and civil society from a broad range of countries that provide AI-specific policy expertise and advice to inform the work of the OECD and GPAI. AI researchers outside the OECD.AI network were also invited to participate in the survey. These individuals are recognised for their active contributions to AI research and policy discussions. They are affiliated with leading institutions[3] and have a median citation count of 2,250,

[3]Their affiliated institutions are the Massachusetts Institute of Technology, Stanford, UC Berkeley, the University of Washington, Princeton, Google DeepMind, Yale, Brown, Korea Advanced Institute of Science and

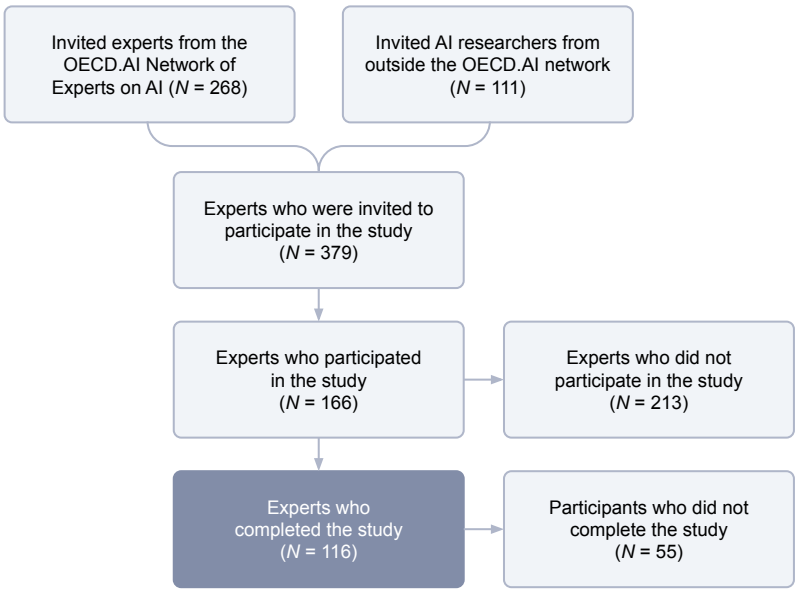

Figure 2: Sample selection process

ranging from a minimum of 110 citations to a maximum of 164,469. Figure 2 illustrates the sample selection process.

Demographics. At the end of the survey, participants were asked to specify their gender (‘woman”, man”, prefer not to say”, and other”), what sector they work in (academia”, civil society”, private sector”, public sector”, and other”), and in what country they live. They were asked to select the option that best describes them. All demographic questions were optional. Table 2 shows the demographic data, including their gender, sector of work, and location.

Gender

Category	Percentage
Men	64.7%
Women	26.7%
Prefer not to say	8.6%

Sector

Category	Percentage
Academia	25.9%
Public sector	24.1%
Private sector	21.6%
Civil society	20.7%
Other	7.7%

Location

Category	Percentage
United States	31.0%
United Kingdom	16.4%
France	13.8%
Canada	5.2%
Others	33.6%

Table 2: Expert survey demographics

Anonymity. Responses to the survey were anonymous by default. At the end of the survey, participants had the option to share their names and affiliations if they wished to be acknowledged in the public write-up of the survey results. A total of 54 participants granted permission to publicly list them as survey respondents; the full list can be found in Appendix A. To protect anonymity and prevent reverse identification, no demographic data or text responses will be made public.

2.1.3 Analysis

Descriptive statistics. For the questions assessing experts’ levels of agreement, only descriptive statistics are reported. These include the percentages of responses for all statements (see Figure 1),

Technology, the Chinese Academy of Sciences, the Center for AI Safety, Northeastern University, the University of Pennsylvania, the Allen Institute for AI, Anthropic, the Collective Intelligence Project, OpenAI, the Vector Institute, McGill University, Trinity College Dublin, the University of Cambridge, the University of Oxford, the Centre for the Governance of AI, HuggingFace, the Signal Foundation, and the Institute for Advanced Study of Princeton University.

Questions used in the expert survey
1. What qualities should thresholds have?
2. How should different types of thresholds be defined?
3. What role should different types of thresholds play?
4. How should thresholds be justified?
5. Who should set thresholds?
6. Who should provide input into the setting of thresholds?
7. How many thresholds should there be?
8. What are potential challenges of setting and evaluating thresholds?
9. When should it be evaluated whether AI systems exceed any thresholds?
10. Who should verify whether AI systems exceed any thresholds?
11. What actions may be warranted if thresholds are exceeded?
12. What mitigations would be adequate?
13. How should thresholds change as AI systems become more capable?

Table 3: Questions used in the expert survey

mean agreements, and standard deviation of mean agreement scores, among other metrics. Due to the relatively small total sample size (N = 166), no additional statistical analyses were performed.

2.2 Method for the public consultation

Purpose. In addition to the expert survey, a public consultation on the OECD.AI website provided an opportunity for anyone to contribute perspectives that may not have been captured by the survey. While the survey questions and statements aimed to be neutral, they inevitably reflected certain viewpoints and potential blind spots. To address these limitations, the public consultation featured more open-ended questions in a less structured format, allowing participants to express views beyond simple agree/disagree responses. Together, these approaches aim to support an analysis that balances inclusivity with technical depth.

Questions. The public consultation included six open-ended questions (see Table 4). To submit a response, participants were required to log in using a service such as Google, Microsoft, or Apple, or sign up for the comment hosting service Disqus. This setup enabled participants to submit free-text responses.

Distribution. The public consultation was open between 26 July 2024 and 1 October 2024. It was announced through a blog post on the OECD.AI website (OECD.AI, 2024) and promoted via email and LinkedIn.

Participants. The consultation was accessible to the public, resulting in 45 participants. Of those, one person submitted anonymously via email. A full list of participants can be found in Appendix B. Notably, three participants of the public consultation also took part in the expert survey.[4]

Coding method. Two policy analysts collaborated to review the comments, identifying themes related to key questions, and developing an initial coding manual focused on topics. Key topics included the benefits and limitations of compute thresholds, types of non-compute thresholds, the complementarity between compute and non-compute thresholds, strategies for identifying and setting thresholds, and government oversight when thresholds are exceeded. Each analyst independently coded the comments using multiple-choice labels for each topic and single-choice labels to indicate the complementarity of compute and non-compute thresholds. They then convened to group similar labels, finalising a

[4]Note that only experts who provided permission for public listing are included (see Appendix A).

Questions used in the public consultation
1. What publications and/or other resources have you found useful on the topic of AI risk thresholds?
2. To what extent do you believe AI risk thresholds based on compute power are appropriate to mitigate risks from advanced AI systems?
3. To what extent do you believe that other types of AI risk thresholds (i.e. thresholds not explicitly tied to compute) would be valuable, and what are they?
4. What strategies and approaches can governments or companies use to identify and set out specific thresholds and measure real-world systems against those thresholds?
5. What requirements should be imposed for systems that exceed any given threshold?
6. What else should the OECD and collaborating organisations keep in mind with regards to designing and/or implementing AI risk thresholds?

Table 4: Questions used in the public consultation

comprehensive list of labels and their descriptions, which can be found in Section 3.2. Finally, they used these labels to assess inter-rater agreement on labelled comments and resolved any remaining discrepancies. The final paper summarises the number of participants who agreed on specific aspects of each topic.

2.3 Method for data triangulation

Purpose. The open-ended responses of the survey and public consultation were combined for analysis. The triangulation approach enables cross-examination of responses from both sources, offering a more nuanced perspective supplementary to notable numerical results and themes from each method. They include tensions relevant to stakeholder involvement in setting thresholds, the appropriateness of thresholds with specific characteristics, the methodological frameworks to justify capability and risk thresholds, and the limitations of thresholds.

Coding method. A manual review of the open-ended responses was conducted independently by three analysts. Responses were retained for further analysis only if at least two out of three reviewers agreed on their relevance to each theme. The majority rule was chosen to balance efficiency and accuracy while processing a large volume of responses. Subsequently, three analysts independently developed mutually exclusive and collectively exhaustive categories to encompass responses relevant to each theme. The research team then conducted a collaborative review to identify common categories. Specific responses were selected and quoted as illustrative examples to supplement and exemplify each theme.

3 Results

This section reports the results from the expert survey (Section 3.1) and public consultation (Section 3.2). Further analysis of open-ended responses can be found below (Section 4).

3.1 Results from the expert survey

Below, selected results across all questions are reported, namely statements with the highest and lowest level of mean agreement (M), the highest proportion of "I don't know" and neither agree nor disagree" responses, notable disagreement between experts, and differences between sectors. The results for all questions and statements are illustrated in Figure 1.

Highest agreement. The five statements with the highest mean agreement (M) across all questions were: "thresholds should be verifiable by external actors" (M = 1.56), "the process of setting thresholds can be influenced by conflicts of interest" (M = 1.54), "if risk thresholds are exceeded,

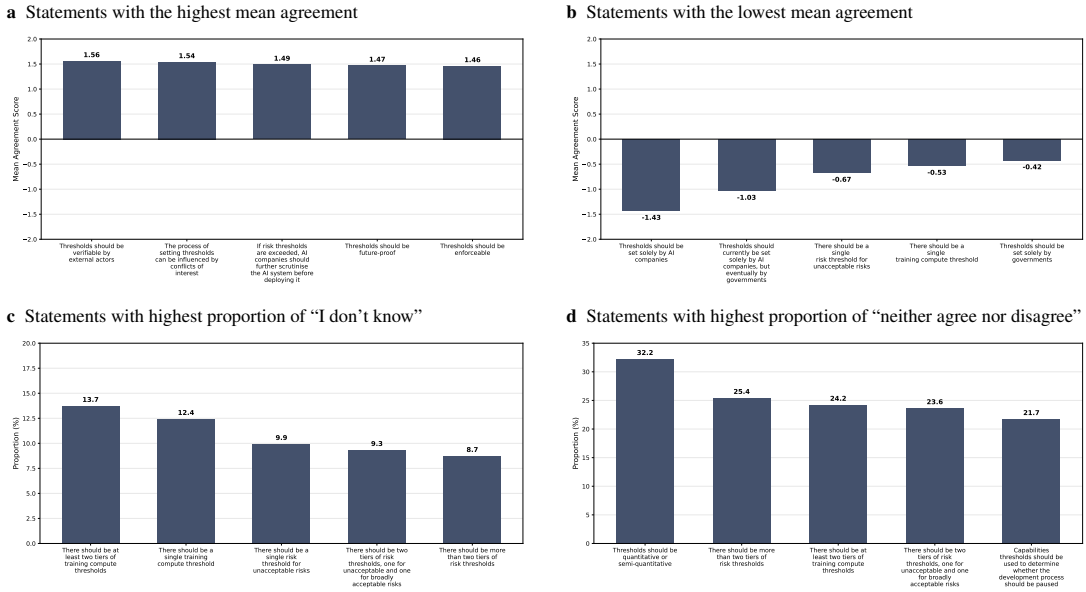

Figure 3: Key results from the expert survey

AI companies should further scrutinise the AI system before deploying it" (M = 1.49), "thresholds should be future-proof" (M = 1.47), and "thresholds should be enforceable" (M = 1.46). Figure 3a presents these five statements with the highest mean agreement.

Lowest agreement. The five statements with the lowest mean agreement (M) across all questions were: "thresholds should be set solely by AI companies" (M = -1.43), "thresholds should currently be set solely by AI companies, but eventually by governments" (M = -1.03), "there should be a single risk threshold for unacceptable risks" (M = -0.67), "there should be a single training compute threshold" (M = -0.53), and "thresholds should be set solely by governments" (M = -0.42). Figure 3b shows the five statements with the lowest mean agreement.

"I don't know". The five statements with the highest proportion of participants responding "I don't know" across all questions were: "there should be at least two tiers of training compute thresholds" (13.7%), "there should be a single training compute threshold" (12.4%), "there should be a single risk threshold for unacceptable risks" (9.9%), "there should be two tiers of risk thresholds, one for unacceptable and one for broadly acceptable risks" (9.3%), and "there should be more than two tiers of risk thresholds" (8.7%). Figure 3c shows the five statements with the highest proportion of "I don't know" responses.

"Neither agree nor disagree". The five statements with the highest proportion of participants responding "neither agree nor disagree" were: "thresholds should be quantitative (i.e. they should be expressed as numerical values) or semi-quantitative (i.e. they should be expressed as ranges of numerical values)" (32.2%), "there should be more than two tiers of risk thresholds" (25.4%), "there should be at least two tiers of training compute thresholds" (24.2%), "there should be two tiers of risk thresholds, one for unacceptable and one for broadly acceptable risks" (23.6%), and "capability thresholds should be used to determine whether the development process should be paused" (21.7%). Figure 3d shows the five statements with the highest proportion of "neither agree nor disagree" responses.

Disagreement between experts. Disagreement among experts quantified using the standard deviation (SD) of mean agreement scores. The five statements with the highest levels of disagreement were: "some thresholds should be defined in terms of training compute" (SD = 0.97), "capability thresholds should be used to determine whether the development process should be paused" (SD = 0.958), "there should be separate thresholds for open models" (SD = 0.945), "the responsibility to identify and implement adequate mitigations should be assigned primarily to AI companies" (SD = 0.919), and "unaccredited or accredited third-party auditors should verify whether AI systems exceed any thresholds" (SD = 0.903).

Differences between private and non-private sectors. Differences between participants from the private sector (*PM*) and non-private sectors (*NM*) are measured by the mean difference (*MD*). Statements were excluded from this ranking if the confidence interval (*CI*) for the mean difference included 0.

For the following statements, participants from the private sector tended to disagree, while participants from non-private sectors tended to agree: "training compute thresholds should be used to identify AI systems that require further scrutiny (e.g. via model evaluations)" (*PM* = -0.167, *NM* = 0.788, *MD* = -0.955, *CI* = -0.148, -0.43) and "there should be separate thresholds for open models, i.e., models with broadly available model weights" (*PM* = -0.417, *NM* = 0.515, *MD* = -0.932, *CI* = -1.45, -0.4).

Participants from the private sector tended to agree less with the following statements than participants from non-private sectors: "if capability thresholds are exceeded, AI companies should notify an independent public body (e.g. the EU AI Office, the Federal Trade Commission [FTC], or an AI Safety Institute)" (*PM* = 0.5, *NM* = 1.333, *MD* = -0.833, *CI* = -1.35, -0.31), "the limitations of existing model evaluations make it challenging to set and evaluate capability thresholds" (*PM* = 0.667, *NM* = 1.455, *MD* = -0.788, *CI* = -1.28, -0.29), and "capability thresholds should be based on risk thresholds such that, if an AI system has certain capabilities, the level of risk will likely be unacceptable" (*PM* = -0.083, *NM* = 0.697, *MD* = -0.78, *CI* = -1.2, -0.36).

3.2 Results from the public consultation

The following section summarises the findings from the public consultation, based on the analysis of responses to the open-ended questions outlined in Table 4. Responses were analysed according the coding scheme described in Section 2.2.

Publications and other resources. Participants recommended 79 resources on thresholds for advanced AI systems. These include academic papers (*N* = 37), blog posts or websites (*N* = 14), government reports, regulation, or guidance (*N* = 13), industry reports (*N* = 12), and other resources (*N* = 2). The most frequently cited resources were: "Risk thresholds for frontier AI" (Koessler et al., 2024) (*N* = 11), "Training compute thresholds: Features and functions in AI regulation" (Heim & Koessler, 2024) (*N* = 8), and "On the limitations of compute thresholds as a governance strategy" (Hooker, 2024) (*N* = 7). The EU AI Act (European Parliament, 2024) (*N* = 5) and the NIST AI Risk Management Framework (NIST, 2023) (*N* = 5) were also mentioned frequently.

Appropriateness of training compute thresholds. Participants were divided on whether training compute thresholds are appropriate to mitigate risks from advanced AI systems. The most commonly mentioned advantage was that, since training compute serves as a proxy for model capabilities, training compute thresholds can be used to evaluate AI systems before development and deployment (*N* = 19). Other advantages mentioned by participants include: training compute thresholds are quantifiable (*N* = 12), enforceable (*N* = 6), simple (*N* = 3), externally verifiable (*N* = 3), objective (*N* = 3), and equitable (*N* = 1). One respondent remarked: "a single, clear, and objective metric makes compute thresholds a strong, first-line regulatory policy lever for mitigating public safety risks from rapid AI changes". Another suggested that compute thresholds as an indicator of computational costs can help policymakers target well-resourced entities while reducing regulatory burdens on smaller ones.

Others raised limitations and concerns. Many participants argued that training compute is an imperfect proxy for risk (*N* = 25). One respondent stated, "compute thresholds alone should generally not determine which mitigation measures are ultimately required, given that compute is only a crude proxy for model capabilities and an even cruder proxy for risks of large-scale societal harm." Participants elaborated on the complex relationship between scale, capabilities, and risks, expressing concerns about reduced oversight (*N* = 4). One participant cautioned that "setting a single number of FLOP could lead to too many models being subjected to additional scrutiny and reporting each year if the threshold is too low. Conversely, if it is set too high, not enough models, including those with existing harms, will be subject to reporting requirements, making the threshold a decorative measure rather than a meaningful indicator of risk."

Additionally, participants noted that the quantitative nature of training compute thresholds can quickly become outdated (*N* = 14) due to advancements in model optimization techniques, existing

inconsistency in the compute upper bound across bills, and susceptibility to Goodhart's Law.[5] One respondent mentioned, "the ease of gameability of compute thresholds raises critical questions about their suitability." Another added, "modern AI techniques increasingly focus on efficiency, enabling high [model] performance with relatively modest computational demands".

Overall, the discussion on compute thresholds revealed varied perspectives. Some believed compute thresholds alone are sufficient and more objective than alternatives ($N = 3$), while others strongly opposed their use ($N = 4$). However, most participants who addressed compute thresholds argued that they should not be used in isolation; instead, they should complement other types of thresholds ($N = 26$).

Value of other types of thresholds. Participants identified various types of threshold. One significant category was capability thresholds, described by model capabilities and adequate mitigations ($N = 19$). These thresholds can inform deployment decisions and aid in developing robust mitigation strategies. Participants emphasised the importance of data quality ($N = 6$), noting that "improvements – such as de-duplication and pruning – can enhance model performance without increasing computational resources".

However, not all suggested thresholds were based on model characteristics. The deployment context and user reach ($N = 14$) also emerged as critical threshold types. One participant suggested that categorising AI systems by their intended use could lead to more tailored regulatory approaches specific to each application area. Participants also discussed risk thresholds based on the probability and severity of harm ($N = 13$). However, some participants expressed concerns about the practicality of using risk-based thresholds, stating that "no one knows or can agree on how to determine the probability." This sentiment underscores the challenges of establishing clear metrics for risk evaluation.

Some participants focused on the autonomy of AI systems ($N = 11$), recommending the identification of potentially dangerous autonomous capabilities, such as autonomous replication or expert-level manipulation abilities. Self-replication was noted as one type of red line ($N = 3$) that indicates unacceptable levels of risk. Economic impact thresholds ($N = 5$) were also suggested, emphasising the economic value added by AI automation in research and development rather than relying solely on quantitative measures.

Other potential metrics for thresholds included the level of transparency ($N = 3$), algorithmic efficiency ($N = 2$), and regulatory compliance ($N = 1$). One participant also highlighted the need to consider the frequency and impact of multi-agent interactions ($N = 1$), as these interactions may lead to emergent behaviours that are difficult to predict during isolated training.

Identifying and setting thresholds. The importance of expert assessment ($N = 12$) and public consultation ($N = 8$) was emphasised by many participants, underscoring the need for collaborative approaches that include diverse stakeholders to avoid regulatory capture and establish standardised metrics. Participants also proposed various strategies for identifying and setting thresholds. Many suggested scenario planning ($N = 10$), which involves "mapping potential risks across various AI-related domains and scenarios."

Several participants highlighted existing AI risk management frameworks from NIST and recommended adopting established safety standards, such as domain-specific regulations like GDPR for data privacy and HIPAA for healthcare applications ($N = 8$). Publications referenced during the public consultation also noted the value of risk assessment frameworks from other safety-critical industries. In sectors like nuclear, maritime, aviation, healthcare, finance, and space, regulators enforce specific risk thresholds (Koessler et al., 2024; Giudici et al., 2023; Cohere for AI, 2024). Safety cases are required in aviation, medical devices, and defence software (Clymer et al., 2024), while scenario analysis is used in national security to enhance threat assessment (Wasil et al., 2024).

Additionally, some participants suggested adaptive testing and monitoring to evaluate capabilities ($N = 4$). Two participants emphasised that companies should bear the responsibility for these assessments through formalised corporate accountability. One participant also called for coordinated global harmonisation and another one for increased investment in research.

Requirements for systems that exceed any threshold. For AI systems exceeding risk thresholds, participants proposed a range of oversight measures. Continuous monitoring with testing was among

[5] Goodhart's Law states that "when a measure becomes a target, it ceases to be a good measure".

the top suggestions (N = 15). Additionally, accountability and liability requirements received strong support (N = 13), with recommendations for structured notifications in case of threshold breaches. One participant suggested that "developers should notify the board of directors, the government, and an independent public body, such as the EU AI Office or the FTC". Another advocated for "public disclosure of safety plans and development roadmaps to promote accountability and allow for external scrutiny of safety measures". Many suggested requiring detailed mitigations measures (N = 10) and mandating third-party evaluation through licensing or certification (N = 5) to ensure that systems adhere to safety standards and address potential risks effectively.

4 Discussion

This section discusses general results that apply to all thresholds (Section 4.1) and specific results that only apply to certain types of thresholds (Section 4.2). It also flags key limitations of the study (Section 4.3) and suggests directions for future research (Section 4.4). Note that all mentions of mean agreements (M) refer to the expert survey, not the public consultation.

4.1 General results

What qualities should thresholds have? Participants largely agreed on the essential qualities thresholds should possess. All relevant statements received high levels of support (see Figure 3a). Participants agreed that thresholds should be verifiable by external actors (M = 1.56), future-proof (M = 1.47),[6] and enforceable by government authorities (M = 1.46).[7] They also supported the need for thresholds to be justified (M = 1.42), account for uncertainties (M = 1.38), and be internationally harmonised (M = 1.14). There was moderate agreement that thresholds should be quantitative or semi-quantitative (M = 0.62).

Participants' support for external verification is also reflected in the literature (Brundage et al., 2020; Avin et al., 2021; Shavit, 2023; Reuel et al., 2024; O'Brien et al., 2024). However, it remains unclear what information external actors need to verify if thresholds are exceeded. This depends on the threshold type, but will likely include information about the AI system (via model or system cards Mitchell et al., 2019; Green et al., 2022), risk assessments (e.g. model evaluation results), mitigations (e.g. cyber defences), and the thresholds themselves.

Efforts to justify different types of thresholds are still nascent. While scholars have suggested options for justifying risk thresholds (Koessler et al., 2024), detailed investigations do not exist. Recent criticisms also highlight the lack of evidence-based justifications for existing training compute thresholds (Hooker, 2024). Although some work on threat modelling exists, much of it remains inaccessible to the public. Additionally, there is often a lack of clear reasoning explaining why AI systems surpassing certain capability thresholds would present significant risks.

Can thresholds from other industries be adopted? Some suggested adopting standards and thresholds from fields such as nuclear, maritime, environment, aviation, and space. However, others warned that AI's unique technical complexity and uncertainty make these standards insufficient (M = 0.82). Thresholds also vary across jurisdictions (Heim & Koessler, 2024).

Who should set thresholds? Participants widely disagreed with statements that thresholds should be set solely by AI companies (M = -1.43), solely by AI companies, with eventual government involvement (M = -1.03), or solely by governments (M = -0.42). Instead, they believed that stakeholder panels, including governments, AI companies, academia, civil society, affected parties, and other relevant stakeholders, should set thresholds (M = 1.08). Participants did not express a strong opinion on whether these thresholds should be set by both governments and AI companies (e.g. governments set the outer bounds and provide high-level guidance, while AI companies set individual thresholds) (M = -0.03).

[6] The term "future-proof" might cause confusion because it can be interpreted in two different ways, namely that thresholds should be robust to technological change or that they should be updated frequently. For the purposes of this study, the latter interpretation was used.

[7] This statement may have been unclear to some respondents. Some may have interpreted it as a statement about the inherent nature of the threshold, while others may have understood it as a normative claim about the legal mechanisms that should be established.

Participants also agreed that thresholds should incorporate input from various sources: academic and research institutions (M = 1.43), civil society organisations (M = 1.28), intergovernmental organisations (M = 1.19), AI companies (M = 1.12), and individuals most affected by the risks (M = 1.04). Input from the public was valued, though to a lesser extent (M = 0.66). Similarly, comments from the public consultation highlighted the need for multiple stakeholders to participate in identifying and setting thresholds, whether that be via expert assessments or public consultation. There is a rich body of literature calling for broad participation (Delgado et al., 2023; Seger, Ovadya, et al., 2023), especially by those affected most by risks, who are often members of historically marginalised communities (Mohamed et al., 2020; Birhane, Isaac, et al., 2022; Birhane, Ruane, et al., 2022). However, some participants also cautioned that broad participation could present its own challenges. Since the public, including people most affected by the risk, often lack sufficient technical knowledge, their ability to provide meaningful input might be limited.

What are the main challenges of setting and evaluating thresholds? Participants expressed strong concern that conflicts of interest can affect the process of setting thresholds (M = 1.54). This score represents the second-highest level of agreement among all questions (see Figure 3a). Additionally, participants recognized that thresholds can be subject to Goodhart's Law (M = 1.20), which states that "when a measure becomes a target, it ceases to be a good measure". Public consultations raised concerns about Goodhart's Law, particularly regarding the compute threshold due to its quantitative nature. For a discussion of threshold-specific challenges, see Section 4.2.

When to evaluate if thresholds have been exceeded. Participants widely agreed that evaluations should occur to determine if AI systems exceed any thresholds before deployment (M = 1.46) and after deployment (M = 1.04). They also supported regular evaluations at set intervals (e.g. every X months or with an increase in training compute by Y) (M = 1.41) and whenever a new version is released (M = 1.30). To a lesser extent, they agreed that evaluations should occur during training (M = 0.83). Participants were neutral about evaluations taking place before training (M = 0.46).

Who should verify if thresholds have been exceeded? Participants strongly agreed that accredited third parties should verify whether AI systems exceed any thresholds (M = 1.28). There was moderate agreement that state actors (M = 0.89) or AI companies themselves (M = 0.79) could perform this verification. Participants were less certain about the role of unaccredited or accredited third-party auditors (M = 0.46).

What actions may be warranted if thresholds are exceeded? Participants agreed that AI companies should publicly disclose any instances of exceeding thresholds (M = 0.81). Many inputs from the public consultation also stressed the need for mandatory reporting (see Kolt et al. (2024)). For a discussion of threshold-specific measures, see Section 4.2.

What mitigations would be adequate? Participants agreed that AI companies should be incentivised to refine existing and develop new mitigations (M = 1.20). Since AI companies employ many AI safety researchers and have more resources than other research institutions, they are in a unique position to contribute to such efforts (Schuett et al., 2024). Participants also agreed that it is still unclear what specific mitigations would be adequate for specific levels of capabilities (M = 1.13). They neither agreed nor disagreed about assigning the primary responsibility for identifying and implementing adequate mitigations to AI companies (M = 0.48).

4.2 Specific results for different types of thresholds

Participants agreed that multiple thresholds should exist, each serving different roles and defined by different metrics (M = 1.13). While they agreed that there should be capability thresholds (M = 1.13), red lines (M = 1.01), and risk thresholds (M = 1.07), they neither agreed nor disagreed with the need for training compute thresholds (M = 0.28). Comments from the public consultation suggested that if training compute thresholds are adopted, complementary non-compute thresholds should also be established.

Below, specific results for different types of thresholds are discussed, namely training compute thresholds (Section 4.2.1), capability thresholds (Section 4.2.2), red lines (Section 4.2.3), risk thresholds (Section 4.2.4) and other thresholds (Section 4.2.5). Values in brackets refer to the mean agreement (M) on a scale from -2 ("strongly disagree") to 2 ("strongly agree").

4.2.1 Training compute thresholds

By "training compute thresholds", we mean thresholds defined in terms of the computational resources used to train a model, often measured in floating point operations (FLOP) (Heim & Koessler, 2024; Hooker, 2024; Koessler et al., 2024; Pistillo et al., 2025).

Training compute thresholds are already part of existing AI regulations and policy initiatives. For example, the EU AI Act uses a threshold of 10^{25} operations to identify general-purpose AI (GPAI) models with systemic risks (European Parliament, 2024; European Commission, 2025b). Similarly, the (rescinded) US Executive Order on Safe, Secure, and Trustworthy AI imposes requirements on companies that train models using more than 10^{26} operations (The White House, 2023). A small, but growing body of literature on training compute thresholds (Heim & Koessler, 2024; Hooker, 2024; Koessler et al., 2024; Sastry et al., 2024; Frontier Model Forum, 2024a; Cottier & Owen, 2025).

What role should training compute thresholds play? Participants widely agreed with the statement that training compute thresholds should not be used alone to determine if an AI system poses unacceptable risks, mainly because training compute is an imperfect proxy for risk ($M = 1.20$). Some survey participants agreed that training compute thresholds could help identify AI systems that require further scrutiny (e.g. via model evaluations) ($M = 0.68$). These findings are consistent with views expressed in the literature (Heim & Koessler, 2024; Hooker, 2024; Koessler et al., 2024) and current practices (European Parliament, 2024; European Commission, 2025b; The White House, 2023). Many public consultation responses echoed concerns about training compute being inadequate to address comprehensive types of risks. Others worried that compute thresholds could reduce oversight for existing models that require less computational power but still pose significant risks.

How to set training compute thresholds? Participants did not express a strong opinion on whether training compute thresholds should be informed by scaling laws ($M = 0.31$). Scaling laws are power laws according to which the use of more data and more compute to train bigger models leads to predictable improvements in their performance (Hestness et al., 2017; Kaplan et al., 2020; Bahri et al., 2021; Hoffmann et al., 2022). However, it remains unclear to what extent current scaling laws will hold in the future (Lohn & Musser, 2022; Villalobos et al., 2024; Sevilla et al., 2024; Narayanan & Kapoor, 2024).

How many training compute thresholds should there be? Participants disagreed that there should be a single training compute threshold ($M = -0.53$). However, they did not express a strong opinion about the statement that there should be at least two training compute thresholds ($M = 0.24$). For both statements, the proportion of participants who responded "I don't know" and "neither agree nor disagree" were among the highest across all questions (see Figure 3c and 3d).

What should companies do if training compute thresholds are exceeded? Participants agreed that if training compute are exceeded, AI companies should conduct additional risk assessments (e.g. via model evaluations) ($M = 0.97$), notify an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute) ($M = 0.79$), and notify the government ($M = 0.65$).

4.2.2 Capability thresholds

By "capability thresholds", we mean thresholds defined in terms of model capabilities and adequate mitigations (Koessler et al., 2024; Frontier Model Forum, 2025c; METR, 2025). Model capabilities are typically assessed via model evaluations (Shevlane et al., 2023; Phuong et al., 2024; Weidinger, Barnhart, et al., 2024; Frontier Model Forum, 2024b), benchmarks (A. K. Zhang et al., 2024; Li et al., 2024; Huang et al., 2024; Laurent et al., 2024; Laine et al., 2024), and red-teaming exercises (Ganguli et al., 2022; Perez et al., 2022; Weidinger, Mellor, et al., 2024), among other things (Frontier Model Forum, 2025a). Mitigations include fine-tuning (Christiano et al., 2017; Ziegler et al., 2019), security controls (Nevo et al., 2024), and access restrictions (O'Brien et al., 2023), among other things (Frontier Model Forum, 2025b; Saeri et al., 2025). A key challenge of setting capability threshold lies in determining what mitigations are adequate for varying levels of capabilities.

What role should capability thresholds play? Participants widely agreed that capability thresholds should indicate when additional mitigations are warranted ($M = 1.21$). They also somewhat agreed that these thresholds could help decide whether to deploy an AI system ($M = 0.79$), roll back or shut down a deployed AI system ($M = 0.71$), or pause the development process ($M = 0.50$). These findings align with the Ministerial Statement (DSIT, 2024b), the Frontier AI Safety Commitments (DSIT, 2024a), and the International Scientific Report on the Safety of Advanced AI (Bengio et al., 2025).

Several AI companies currently use capability thresholds to assess whether to pause development processes (Anthropic, 2025; OpenAI, 2025; Google DeepMind, 2025).

How to set capability thresholds? Participants agreed that capability thresholds should be informed by threat models ($M = 1.05$), which describe how different risk factors might lead to harm. They also agreed to a lesser extent that these thresholds should be guided by human uplift studies ($M = 0.61$). Human uplift studies assess how access to a specific AI system improves human performance. Additionally, participants agreed that capability thresholds should align with risk thresholds, indicating that certain capabilities likely lead to unacceptable risk ($M = 0.54$). Recent literature supports this argument (Koessler et al., 2024).

However, participants expressed concern over the limitations of existing model evaluations, which complicate the setting and evaluation of capability thresholds ($M = 1.21$). This concern was more prevalent in non-private sectors. Recent research highlights limitations in current model evaluations (Gehrmann et al., 2022; Anthropic, 2023; van der Weij et al., 2024; Järviniemi & Hubinger, 2024; Kapoor, Stroebl, et al., 2024; Rauh et al., 2024). Participants also noted that the lack of agreed-upon threat models for many risks and the diversity of threat models for general-purpose AI systems hinder setting capability thresholds ($M = 1.06$).

How many capability thresholds should there be? Participants agreed that multiple tiers of capability thresholds ($M = 0.90$) are necessary, as well as different thresholds for different types of capabilities ($M = 1.06$).

What should companies do if capability thresholds are exceeded? Participants widely agreed that, if capability thresholds are exceeded, AI companies should implement additional mitigations ($M = 1.23$). They also agreed that AI companies should notify key stakeholders, namely their board of directors ($M = 1.20$), an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute) ($M = 1.16$), and the government ($M = 1.04$). To a lesser extent, they agreed that AI companies should prepare a safety case if thresholds are exceeded ($M = 0.97$). A safety case is a structured argument, supported by evidence, that a system is sufficiently safe (Clymer et al., 2024; Buhl et al., 2024; Goemans et al., 2025). Participants also agreed that if capability thresholds are exceeded and AI companies cannot implement adequate mitigations, AI companies should pause the development and deployment process until the thresholds are no longer exceeded ($M = 0.83$). Existing safety frameworks, including Anthropic's Responsible Scaling Policy (Anthropic, 2025), OpenAI's Preparedness Framework (OpenAI, 2025), and Google DeepMind's Frontier Safety Framework (Google DeepMind, 2025), already contain commitments along these lines (METR, 2025). This practice has also been supported in the literature (Alaga & Schuett, 2023).

4.2.3 Red lines

In this paper, "red lines" refer to thresholds based on unacceptable model capabilities, regardless of any mitigations. It is important to note that there is no universally accepted definition of "red lines". The most relevant source on red lines is a consensus statement from the International Dialogues on AI Safety (IDAIS) in March 2024, which outlines proposed red lines for advanced AI systems from Western and Chinese scientists (IDAIS, 2024). However, literature on the subject remains limited (Bengio et al., 2024; Zoumpalova & Iliadis, 2025).

What role should red lines play? Participants widely agreed that red lines should be used to determine whether the development process should be paused ($M = 0.93$) or when a deployed AI should be rolled back or shut down ($M = 1.07$). As previously mentioned, current safety frameworks already include commitment for capability thresholds (see Section 4.2.2). Various roll-back and shut-down mechanisms have also been explored in the literature (O'Brien et al., 2024).

How many red lines should there be? Participants concurred that there should be different red lines for different types of capabilities ($M = 0.97$). It is essential to clarify that there can only be one tier of red lines. In a recent consensus statement (IDAIS, 2024), scientists propose a non-exhaustive list of red lines for advanced AI systems (see Table 5).

What should companies do if red lines are crossed? Participants overwhelmingly agreed that, if red lines are crossed, AI companies should notify their board of directors ($M = 1.38$), an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute) ($M = 1.35$), and the government ($M = 1.26$). Additionally, AI companies should pause the development and deployment process until the red lines are no longer exceeded ($M = 1.27$).

Red line	Description
Autonomous replication or improvement	No AI system should be able to copy or improve itself without explicit human approval and assistance. Examples include both exact copies of itself as well as creating new AI systems of similar or greater abilities.
Power seeking	No AI system should take actions to unduly increase its power and influence.
Assisting weapon development	No AI systems should substantially increase the ability of actors to design weapons of mass destruction, or violate the biological or chemical weapons convention.
Cyberattacks	No AI system should be able to autonomously execute cyberattacks resulting in serious financial losses or equivalent harm.
Deception	No AI system should be able to consistently cause its designers or regulators to misunderstand its likelihood or capability to cross any of the preceding red lines.

Table 5: Red lines for advanced AI systems

4.2.4 Risk thresholds

By "risk thresholds", we mean thresholds defined in terms of the probability and severity of harm (Koessler et al., 2024), corresponding with the "risk" definition found in the EU AI Act (European Parliament, 2024) and NIST AI Risk Management Framework (NIST, 2023).

Risk thresholds are common in other safety-critical industries, such as nuclear, maritime, aviation, and space. They are also integral to AI risk management standards, such as the NIST AI Risk Management Framework (NIST, 2023) and corresponding guidelines (NIST, 2024) as well as ISO/IEC 23894 (ISO & IEC, 2023). While extensive literature exists on risk thresholds in other sectors (Linkov et al., 2011; Marhavilas & Koulouriotis, 2021; Fischhoff et al., 1981; Klinke & Renn, 2002; Starr, 1969), literature on AI risk thresholds is still nascent (Koessler et al., 2024; Raman et al., 2025; Caputo et al., 2025).

What role should risk thresholds play? Participants agreed that risk thresholds should inform deployment decisions (M = 1.22). To a lesser extent they also supported using risk thresholds to set capability thresholds (M = 0.56). The literature has suggested a distinction between "direct" and "indirect" ways in which risk thresholds can inform deployment decisions which roughly correspond to these two options (Koessler et al., 2024).

How to set risk thresholds? Participants expressed equal agreement on how to set risk thresholds based on three criteria: expert opinions (M = 0.63), existing thresholds from other safety-critical industries (M = 0.61), and cost-benefit analyses (M = 0.59). They neither agreed nor disagreed that risk thresholds should reflect people's revealed preferences regarding acceptable risk levels in everyday activities (M = -0.02). "Revealed preferences" refer to the level of risk people seem to accept when engaging in common activities (e.g. driving). All four approaches have been used in other industries (Koessler et al., 2024).

Concerns. Concerns were raised about the potential for AI companies to downplay risks to remain within acceptable thresholds (M = 1.44) and the lack of reliable risk estimation methods complicating the setting and evaluation of these thresholds (M = 1.23). These concerns were also voiced by participants of the public consultation. Additionally, there were reservations about the applicability of existing thresholds from other industries to advanced AI systems, which pose unique challenges (M = 0.82).

How many risk thresholds should there be? Most participants agreed that multiple risk thresholds should exist. They supported having a threshold above which risks are considered unacceptable (M = 1.25) and another below which risks are broadly acceptable, requiring no additional mitigations (M = 0.91). They also believed that, between the thresholds for broadly acceptable and unacceptable risk, AI companies should implement additional mitigations unless their costs are grossly disproportionate

to their benefits ($M = 0.90$), a principle known "as low as reasonably practicable (ALARP)" (Melchers, 2001; Linkov et al., 2011). Overall, participants disagreed with the notion of a single risk threshold for unacceptable risks ($M = -0.67$), but showed uncertainty regarding the ideal number of thresholds, neither agreeing nor disagreeing that there should be two ($M = 0.02$) and providing limited support for more than two tiers of risk thresholds ($M = 0.50$). Importantly, they believed that there should be different risk thresholds for different types of harms (e.g. number of fatalities and economic damage) ($M = 1.22$).

What should companies do if risk thresholds are exceeded? Participants generally agreed that exceeding risk thresholds should prompt AI companies to further scrutinise the AI system before deployment (e.g. through model evaluations) ($M = 1.49$). They did not express a strong opinion on whether exceeding risk thresholds should lead to a strict decision rule (e.g. "if an AI system exceeds a risk threshold, it may not be deployed"), primarily due to the unreliability of current risk estimation methods ($M = 0.00$).

4.2.5 Other types of thresholds

Participants expressed agreement that there should be additional types of thresholds ($M = 0.94$). Feedback from the public consultation suggested that these thresholds could be defined by various factors, including, but not limited to, data quality, algorithmic efficiency, interaction effects of multiple agents, the number and type of users, deployment context, and release strategy. The term "deployment context" refers to the way in which an AI system is deployed. While some systems are deployed via an application programming interface (API) (Shevlane, 2022; Bucknall & Trager, 2023), others are open-sourced (Seger, Dreksler, et al., 2023; Kapoor, Bommasani, et al., 2024; Bommasani et al., 2024). But note that there is a whole spectrum of deployment modes (Solaiman, 2023). The Frontier Safety Commitments are also open to different types of thresholds (DSIT, 2024a).

4.3 Limitations

The results should be seen as suggestive rather than authoritative, as the study may not capture the full range of views across the community of AI experts and the public for the following reasons.

Survey sample. The expert survey, designed to explore a wide range of claims across 98 aspects of setting, verifying, and enforcing thresholds, was limited by a small sample size ($N = 166$). In addition, the number of responses varied below 166 for most questions as they were optional which may affect the robustness of the findings. See Figure for the number of responses for each claim. Additionally, the sample did not achieve a balanced representation of private and non-private sectors, potentially influencing the assessment of dissensus between these groups (see Table 2).

Terms. While key terms were defined at the outset of the expert survey (see Table 1) to provide clarity, these terms are not universally accepted and may have been interpreted differently by participants. The lack of a common understanding regarding terms relevant to thresholds may have distorted the results to some degree. For example, some individuals and organisations use "risk threshold" and "capability threshold" interchangeably. OpenAI's Preparedness Framework refers to "risk thresholds" (OpenAI, 2025), whereas other sources use the term "capability thresholds" (Koessler et al., 2024). Notably, the initial OECD.AI blog post announcing this project also mentioned "risk thresholds" (OECD.AI, 2024), although it encompassed a broader scope, including various types of thresholds, such as training compute thresholds.

Survey questions. Although the expert survey was crafted to be neutral and refined through feedback from partner research organisations, it inevitably reflects certain perspectives and assumptions. Some participants pointed out that the survey's focus on thresholds for AI models may restrict discussions related to risks associated with larger AI systems or the interactions between multiple AI systems. A few respondents noted that the survey appeared to assume a direct correlation between increasing AI capabilities and heightened risks. Moreover, some survey statements could have been clearer to help participants better evaluate them. Phrases like "thresholds should be future-proof", "thresholds should be enforceable", or "risk thresholds should be used to set capability thresholds" may have been ambiguous, leading to varied interpretations. Four participants thought that the survey design was not suitable for encompassing diverse viewpoints. They thought the fixed response format was restrictive for those who disagreed with the underlying assumptions and terminologies. However,

these concerns are at least partially addressed by the open-ended response options in the survey and the public consultation, along with triangularisation during analysis.

Coding. Analysis of open-ended responses was conducted by three analysts, following best practices for qualitative coding, including a structured coding framework and regular discussions to ensure intercoder reliability. Despite these efforts, some subjectivity may still have been introduced.

4.4 Future work

The study raises several unanswered questions, underscoring the urgent need for further research in this area. Future efforts to establish thresholds for AI systems should focus on defining and operationalizing these thresholds in relation to model capabilities and key risk mitigations, particularly through the development of comprehensive threat models. A structured process for creating stakeholder panels may be essential to facilitate inclusive decision-making, ensuring that diverse perspectives from government, industry, and civil society are represented. Additionally, identifying mechanisms for external verification by accredited third parties could enhance the credibility and reliability of threshold assessments. Finally, addressing potential conflicts of interest in the threshold-setting process will be important to ensure balanced and equitable input from all stakeholders involved. By pursuing these avenues, the field can enhance the robustness and effectiveness of thresholds as governance tools for AI systems.

5 Conclusion

This paper has reported findings from the first survey on thresholds for advanced AI systems. It has identified several areas of agreement and divergence among experts from academia, civil society, as well as the private and public sector. The findings can serve as evidence in ongoing policy discussions. In light of rapid progress in AI development, setting thresholds for advanced AI systems is an urgent and difficult challenge that governments and AI companies cannot – and should not – tackle on their own. Instead, different stakeholder groups must work together and more research is needed to address some of the many open problems. The authors hope that this study can contribute to such efforts.

Acknowledgments

We would like to thank all experts who filled out the survey and everyone who participated in the public consultation. We are grateful for valuable feedback, suggestions, and support from Ben Garfinkel, Benjamin Prudhomme, Charbel Segerie, Charbel-Raphael Segerie, Clara Neppel, Conrad S. Tuck, Daniel Kossack, Daniel Privitera, Hamish Hobbs, Henry Papadatos, Hiroki Habuka, James Gealy, Jesse Dunietz, Johannes Jutting, Lennart Heim, Leonie Koessler, Markus Anderljung, Noemi Dreksler, Raja Chatila, Sebastian Hallensleben, Shayne Longpre, Siméon Campos, Stuart Russell, Taylor Reynolds, Tegan McCaslin, and Yoshua Bengio (in alphabetical order).

References

Alaga, J., & Schuett, J. (2023). Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers. *arXiv preprint arXiv:2310.00374*.

Anthropic. (2023). *Challenges in evaluating AI systems.* https://www.anthropic.com/research/evaluating-ai-systems.

Anthropic. (2025). *Responsible Scaling Policy.* https://anthropic.com/rsp.

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., … Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science*, *374*(6573), 1327–1329. https://doi.org/10.1126/science.abi7176.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2021). Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *384*(6698), 842–845. https://doi.org/10.1126/science.adn0117.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... Dietterich, T. G. (2025). *International AI safety report.* https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–8). https://doi.org/10.1145/3551624.3555290.

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., & L. Dancy, C. (2022). The forgotten margins of AI ethics. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 948–958). https://doi.org/10.1145/3531146.3533157.

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., ... Liang, P. (2024). Considerations for governing open foundation models. *Science*, *386*(6718), 133–136. https://doi.org/10.1126/science.adp1848.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Bucknall, B. S., & Trager, R. F. (2023). *Structured access for third-party research on frontier AI models.* Oxford Martin AI Governance Initiative. https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf.

Buhl, M. D., Schuett, J., & Anderljung, M. (2024). Safety cases for frontier AI. *arXiv preprint arXiv:2410.21572*.

Caputo, N. A., Campos, S., Casper, S., Gealy, J., Hung, B., Jacobs, J., ... Trager, R. (2025). *Risk tiers: Towards a gold standard for advanced AI.* Oxford Martin AI Governance Initiative. https://aigi.ox.ac.uk/publications/risk-tiers-towards-a-gold-standard-for-advanced-ai.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). Safety cases: How to justify the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462*.

Cohere for AI. (2024). *The limits of thresholds.* https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf.

Cottier, B., & Owen, D. (2025). *How many AI models will exceed compute thresholds?* Epoch. https://epoch.ai/blog/model-counts-compute-thresholds.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in AI design: Theoretical foundations and the current state of practice. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–23). https://doi.org/10.1145/3617694.3623261.

DSIT. (2023). *The Bletchley Declaration by countries attending the AI Safety Summit.* https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

DSIT. (2024a). *Frontier AI Safety Commitments, AI Seoul Summit 2024.* https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024.

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

European Commission. (2025a). *The General-Purpose AI Code of Practice.* https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

European Commission. (2025b). *Guidelines on the scope of obligations of providers of general-purpose AI models under the AI Act.* https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act.

European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1981). *Acceptable risk*. Cambridge University Press.

Frontier Model Forum. (2024a). *Measuring training compute.* https://www.frontiermodelforum.org/updates/issue-brief-measuring-training-compute.

Frontier Model Forum. (2024b). *Preliminary taxonomy of pre-deployment frontier AI safety evaluations.* https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations.

Frontier Model Forum. (2025a). *Frontier capability assessments.* https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments.

Frontier Model Forum. (2025b). *Frontier mitigations.* https://www.frontiermodelforum.org/technical-reports/frontier-mitigations.

Frontier Model Forum. (2025c). *Risk taxonomy and thresholds.* https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds.

G7. (2023). *Hiroshima process international guiding principles for organizations developing advanced AI system.* https://www.mofa.go.jp/files/100573471.pdf.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gehrmann, S., Clark, E., & Sellam, T. (2022). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Giudici, P., Centurelli, M., & Turchetta, S. (2023). Artificial intelligence risk measurement. *Expert Systems with Applications*, *235*, 121220. https://doi.org/10.1016/j.eswa.2023.121220.

Goemans, A., Buhl, M., Schuett, J., Korbak, T., Wang, J., Hilton, B., & Irving, G. (2025). Safety case template for frontier AI: A cyber inability argument. *arXiv preprint arXiv:2411.08088*.

Google DeepMind. (2025). *Updating the Frontier Safety Framework.* https://deepmind.google/discover/blog/updating-the-frontier-safety-framework.

Green, N., Procope, C., Cheema, A., & Adediji, A. (2022). *System cards: A new resource for understanding how AI systems work.* Meta. https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work.

Heim, L., & Koessler, L. (2024). Training compute thresholds: Features and functions in AI regulation. *arXiv preprint arXiv:2405.10799*.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hooker, S. (2024). On the limitations of compute thresholds as a governance strategy. *arXiv preprint arXiv:2407.05694*.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., ... Zhao, Y. (2024). TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

IDAIS. (2024). *IDAIS-Beijing, 2024.* https://idais.ai/dialogue/idais-beijing.

ISO, & IEC. (2023). *Information technology – Artificial intelligence – Guidance on risk management.* https://www.iso.org/standard/77304.html.

Järviniemi, O., & Hubinger, E. (2024). Uncovering deceptive tendencies in language models: A simulated company AI assistant. *arXiv preprint arXiv:2405.01576*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... Narayanan, A. (2024). On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI agents that matter. *arXiv preprint arXiv:2407.01502*.

Klinke, A., & Renn, O. (2002). A new approach to risk evaluation and management: Risk-based, precaution-based, and discourse-based strategies. *Risk Analysis*, *22*(6), 1071–1094. https://doi.org/10.1111/1539-6924.00274.

Koessler, L., Schuett, J., & Anderljung, M. (2024). Risk thresholds for frontier AI. *arXiv preprint arXiv:2406.14713*.

Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., ... Woodside, T. (2024). Responsible reporting for frontier AI development. *arXiv preprint arXiv:2404.02675*.

Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., ... Evans, O. (2024). Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. *arXiv preprint arXiv:2407.04694*.

Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., ... Rodriques, S. G. (2024). LAB-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., ... Hendrycks, D. (2024). The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Linkov, I., Bates, M., Loney, D., Sparrevik, M., & Bridges, T. (2011). Risk management practices. In I. Linkov & T. Bridges (Eds.), *Climate: Global change and local adaptation* (pp. 133–155). Springer. https://doi.org/10.1007/978-94-007-1770-1_8.

Lohn, A., & Musser, M. (2022). *AI and compute: How much longer can computing power drive artificial intelligence progress?* Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA009.

Marhavilas, P. K., & Koulouriotis, D. E. (2021). Risk-acceptance criteria in occupational health and safety risk-assessment: The state-of-the-art through a systematic literature review. *Safety*, *7*(4), 77. https://doi.org/10.3390/safety7040077.

Melchers, R. E. (2001). On the ALARP approach to risk management. *Reliability Engineering & System Safety*, *71*(2), 201–208. https://doi.org/10.1016/S0951-8320(00)00096-X.

METR. (2025). *Common elements of frontier AI safety policies.* https://metr.org/common-elements.pdf.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … Gebru, T. (2019). Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 220–229). https://doi.org/10.1145/3287560.3287596.

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8.

Narayanan, A., & Kapoor, S. (2024). *AI scaling myths.* AI Snake Oil. https://www.aisnakeoil.com/p/ai-scaling-myths.

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (2024). *Securing AI model weights: Preventing theft and misuse of frontier models.* RAND. https://doi.org/10.7249/RRA2849-1.

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* https://doi.org/10.6028/NIST.AI.100-1.

NIST. (2024). *Managing misuse risk for dual-use foundation models: Initial public draft.* https://doi.org/10.6028/NIST.AI.800-1.ipd.

O'Brien, J., Ee, S., Kraprayoon, J., Anderson-Samways, B., Delaney, O., & Williams, Z. (2024). Coordinated disclosure of dual-use capabilities: An early warning system for advanced AI. *arXiv preprint arXiv:2407.01420*.

O'Brien, J., Ee, S., & Williams, Z. (2023). Deployment corrections: An incident response framework for frontier AI models. *arXiv preprint arXiv:2310.00328*.

OECD.AI. (2024). *Public consultation on risk thresholds for advanced AI systems.* https://oecd.ai/en/wonk/seeking-your-views-public-consultation-on-risk-thresholds-for-advanced-ai-systems-deadline-10-september.

OpenAI. (2025). *Updating our Preparedness Framework.* https://openai.com/index/updating-our-preparedness-framework.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., … Irving, G. (2022). Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., … Shevlane, T. (2024). Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*.

Pistillo, M., Arsdale, S. V., Heim, L., & Winter, C. (2025). The role of compute thresholds for AI governance. *George Washington Journal of Law & Technology.*, *1*(1), 1–25.

Raman, D., Madkour, N., Murphy, E. R., Jackson, K., & Newman, J. (2025). Intolerable risk threshold recommendations for artificial intelligence. *arXiv preprint arXiv:2503.05812*.

Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Comanescu, R., Akbulut, C., … Weidinger, L. (2024). Gaps in safety evaluations of generative AI. In *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 1039–1052). https://doi.org/10.1609/aies.v7i1.31717.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., … Trager, R. (2024). Open problems in technical AI governance. *arXiv preprint arXiv:2407.14981*.

Saeri, A., Graham, S. L. G. J., Lacarriere, C., Slattery, P., & Thompson, N. (2025). *Mapping AI risk mitigations.* MIT AI Risk Repository. https://airisk.mit.edu/blog/mapping-ai-risk-mitigations.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., … Coyle, D. (2024). Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*.

Schuett, J., Anderljung, M., Carlier, A., Koessler, L., & Garfinkel, B. (2024). From principles to rules: A regulatory approach for frontier AI. *arXiv preprint arXiv:2407.07300*.

Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023). Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., … Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple meanings, goals, and methods. In *ACM Conference on AI, Ethics, and Society* (pp. 715–722). https://doi.org/10.1145/3600211.3604693.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., & Erdil, E. (2024). *Can AI scaling continue through 2030?* Epoch. https://epochai.org/blog/can-ai-scaling-continue-through-2030.

Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*.

Shevlane, T. (2022). Structured access: An emerging paradigm for safe AI deployment. In *The oxford handbook of ai governance.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.39.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., … Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. *arXiv preprint arXiv:2302.04844*.

Starr, C. (1969). Social benefit versus technological risk: What is our society willing to pay for safety? *Science*, *165*(3899), 1232–1238. https://doi.org/10.1126/science.165.3899.1232.

The White House. (2023). *Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*.

Wasil, A., Smith, E., Katzke, C., & Bullock, J. (2024). AI emergency preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats. *arXiv preprint arXiv:2407.17347*.

Weidinger, L., Barnhart, J., Brennan, J., Butterfield, C., Young, S., Hawkins, W., … Isaac, W. (2024). Holistic safety and responsibility evaluations of advanced AI models. *arXiv preprint arXiv:2404.14068*.

Weidinger, L., Mellor, J., Pegueroles, B. G., Marchal, N., Kumar, R., Lum, K., … Isaac, W. (2024). STAR: Sociotechnical approach to red teaming language models. *arXiv preprint arXiv:2406.11757*.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., … Liang, P. (2024). Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research*, *71*, 591–666. https://doi.org/10.1613/jair.1.12895.

Zhang, B., & Dafoe, A. (2020). U.S. public opinion on the governance of artificial intelligence. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–193. https://doi.org/10.1145/3375627.3375827.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., … Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Zoumpalova, T., & Iliadis, A. (2025). *AI red lines.* The Future Society. https://thefuturesociety.org/airedlines.

Appendix

Appendix A: List of survey participants

54 participants of the expert survey consented to their names and affiliations, as specified by them, being mentioned in this paper (in alphabetical order). 111 participants, not listed here, did not provide their permission. Note that participants do not necessarily represent any organisations they are affiliated with. They chose to add their name after completing the survey and were not sent the manuscript before publication.

- Adriano Koshiyama, Holistic AI
- Amit Ashkenazi
- Armando Guio, Global Network of Internet & Society Centers (NoC)
- Amanda Craig, Owen Larter, Ani Gevorkian & William Bartholomew Microsoft
- Ben Bucknall, Centre for the Governance of AI
- Benoit Bergeret, strategies.ai
- Blanc Nicolas, TUAC
- Brendan Reidenbach, International Energy Agency
- Carlos I Gutierrez, Google
- Carlos Muñoz Ferrandis
- Charles Fadel, BIAC
- Charles Martinet & Charbel-Raphaël Segerie, Centre pour la Sécurité de l'IA (CeSIA)
- Craig E. Shank, CES.WORLD PLLC
- Cyrus Hodes
- Derli Anacona, Departamento Nacional de Planeación
- Dexter Docherty, OECD
- Eric Sutherland, OECD
- Eva Thelisson, AI Transparency Institute
- Evan Hadfield, The Collective Intelligence Project
- Graham Taylor, University of Guelph / Vector Institute
- Gregg Barrett, Cirrus AI
- Henry Papadatos, SaferAI
- Holden Karnofsky, Carnegie Endowment for International Peace
- Ian R. Hodgkinson, Loughborough Business School, Loughborough University, UK
- Ilana Golbin Blumenfeld, PwC
- Kate Kaye, World Privacy Forum
- Kwak Joon-ho, TTA of the ROK
- Lennart Heim, RAND
- Liliana Fernandez Gomez, Spiral Center of Technologies for Development
- Luis Ricardo Sánches Hernandez, National Institute for Transparency Access to Public Information and Personal Data Protection
- Markus Anderljung, Centre for the Governance of AI
- Merve Hickok, Center for AI and Digital Policy
- Michel Morvan, Cosmo Tech
- Nico Miailhe, PRISM Eval
- Niloofar Mireshghallah, UW
- Olivia J. Erdelyi, University of Canterbury & University of Bonn

- Ott Velsberg, Ministry of Economic Affairs and Communications
- Qinghua Lu, CSIRO, Australia
- Rachel Freedman, UC Berkeley
- Rafael Cuervo, National Planning Department
- Rebecca Finlay, Partnership on AI
- Richard Mallah, Future of Life Institute
- Samo Zorc, Chair of AI Technical Committee, Slovenian Institute for Standardization (SIST)
- Sean McGregor, UL Research Institutes
- Sebastian Hallensleben, VDE / CEN-CENELEC
- Stephen Casper, MIT
- Tim Clement-Jones, UK House of Lords
- Tim Fist, Institute for Progress
- Tim G. J. Rudner, New York University
- Tom David, PRISM
- Tom Jackson, Loughborough University
- Utpal Mangla, IBM
- Yannis Assael, Ministry of Digital Governance, Greece
- Yeong Zee Kin, Chief Executive, Singapore Academy of Law
- Yoshua Bengio, University of Montreal & Mila

Appendix B: List of consultation participants

The following people participated in the public consultation (in alphabetical order). Note that participants do not necessarily represent any organisations they are affiliated with.

- Ajay Gambhir
- Anna Katariina Wisakanto
- Benjamin Barber
- Brian Scarpelli
- Center for AI and Digital Policy (CAIDP)
- Charbel-Raphael Segerie
- Dave Lewis
- Demetrius Floudas
- Eva Behrens
- Francesca Rossi
- Geetika
- Giacomo Petrillo
- Heather Domin
- Herp Derpingson
- Ima Bello
- James Norris
- John Handy Bosma
- John Sotiropoulos
- Jose Oyola
- Juho Reivo
- Kadian Davis-Owusu
- Kedharnath Sankararaman
- Kiyomi Carbone
- Kyrtin Atreides
- Lennart Heim
- Lilian Do Khac
- Luciano Zorzin
- Majiuzu Daniel Moses
- Mario Bertorelli
- Matteo
- Melissa Hopkins
- Michael Borelli
- Michael Chen
- Nell Watson
- Peter Slattery
- Raja Sengupta
- Rebecca Portnoff
- Sara Hooker
- Seth Hays
- Simon Falk

- Tereza Zoumpalova
- Tom David
- Will Jennings
- Yoshua Bengio
- Zach Stein-Perlman
- Anon1
- Anon2

Appendix C: Questionnaire

1. What qualities should thresholds have?

- "Thresholds should be justified, i.e. they should rest on explicit arguments for why AI systems that exceed the thresholds would pose severe risks."
- "Thresholds should be easy to communicate."
- "Thresholds should be quantitative (i.e. they should be expressed as numerical values) or semi-quantitative (i.e. they should be expressed as ranges of numerical values)."
- "Thresholds should account for uncertainties, i.e. they should reflect the fact that risk assessment methods might be imprecise (e.g. by adding safety margins)."
- "Thresholds should be internationally harmonised, i.e. thresholds in different countries should, as much as possible, be set in similar ways."
- "Thresholds should be verifiable by external actors, i.e. external actors should be provided necessary information to verify whether the thresholds have been exceeded."
- "Thresholds should be future-proof, i.e. they should be updated periodically according to the state of the art."
- "Thresholds should be enforceable, i.e. government authorities should be able to enforce the actions that AI companies should take if thresholds are exceeded."

2. How should different types of thresholds be defined?

- "Some thresholds should be defined in terms of risk estimates, i.e. they should specify what likelihood and magnitude of different types of harm would be acceptable (risk thresholds)."
- "Some thresholds should be defined in terms of model capabilities AND mitigations, i.e. they should specify what capabilities would be concerning and what mitigations would be adequate for those capabilities (capabilities thresholds)."
- "Some thresholds should be defined in terms of unacceptable model capabilities, i.e. they should specify what capabilities would be unacceptable (red lines) regardless of mitigation measures."
- "Some thresholds should be defined in terms of training compute, i.e. the computational resources used to train a model (training compute thresholds)."
- "Some thresholds should be defined in terms of other factors (e.g. number and type of users, deployment context, or release strategy)."
- "There should be different thresholds that play different roles and are defined using different metrics."

3. What role should different types of thresholds play?

- "Risk thresholds should be used to set capabilities thresholds."
- "Risk thresholds should be used to inform the decision about whether an AI system should be deployed."
- "Capabilities thresholds should be used to determine whether additional mitigations are warranted."
- "Capabilities thresholds should be used to determine whether the development process should be paused."
- "Capabilities thresholds should be used to determine whether an AI system should be deployed."
- "Capabilities thresholds should be used to determine whether a deployed AI system should be rolled back or shut down."
- "Red lines should be used to determine whether the development process should be paused."
- "Red lines should be used to determine whether a deployed AI system should be rolled back or shut down."
- "Training compute thresholds should be used to identify AI systems that require further scrutiny (e.g. via model evaluations)."

- "Since training compute is an imperfect proxy for risk, training compute thresholds should not be used on their own to determine whether an AI system poses unacceptable risks."

4. How should thresholds be justified?

- "Risk thresholds should be based on cost-benefit analyses or similar approaches, i.e. analyses that weigh the potential harms and benefits of AI systems."
- "Risk thresholds should be based on existing risk thresholds from other industries, i.e. reviews of risk thresholds in other safety-critical industries (e.g. nuclear, chemicals, or aviation) adapted to AI systems."
- "Risk thresholds should be based on expert opinions, i.e. surveys of the level of risk that a diverse set of experts considers to be acceptable."
- "Risk thresholds should be based on people's revealed preferences, i.e. reviews of the level of risk people seem to accept when engaging in common activities (e.g. driving)."
- "Capabilities thresholds should be based on risk thresholds such that, if an AI system has certain capabilities, the level of risk will likely be unacceptable."
- "Capabilities thresholds should be informed empirical studies that assess how much access to a specific AI system improves human performance (human uplift studies)."
- "Capabilities thresholds should be informed by threat models, i.e. models that describe how different risk factors could plausibly lead to harm."
- "Training compute thresholds should be based on the amount of computational resources used to train existing models."
- "Training compute thresholds should be informed by scaling laws, i.e. power laws according to which the use of more data and more compute to train bigger models leads to predictable improvements in their performance."

5. Who should set thresholds?

- "Thresholds should be set solely by governments."
- "Thresholds should be set solely by AI companies."
- "Thresholds should be set by governments AND AI companies (e.g. governments set the outer bounds and provide high-level guidance, while AI companies set individual thresholds)."
- "Thresholds should be set by stakeholder panels (e.g. composed of governments, AI companies, academia, civil society, affected parties and other relevant stakeholders)."
- "Thresholds should currently be set solely by AI companies, but eventually by governments."

6. Who should provide input into the setting of thresholds?

- "Thresholds should be set with input from AI companies (assuming that thresholds are not set by AI companies)."
- "Thresholds should be set with input from academic and research institutions."
- "Thresholds should be set with input from civil society organisations."
- "Thresholds should be set with input from intergovernmental organisations."
- "Thresholds should be set with input from the public."
- "Thresholds should be set with input from the people most affected by the risks."

7. How many thresholds should there be?

- "There should be a risk threshold above which the level of risk is unacceptable."
- "There should be a risk threshold below which the level of risk is broadly acceptable, i.e. no additional mitigations are needed."
- "Between the thresholds for broadly acceptable and unacceptable risk, AI companies should keep the level of risk as low as reasonably practicable (ALARP), i.e. companies need to implement additional mitigations unless their costs are grossly disproportionate to their benefits."

- "There should be a single risk threshold for unacceptable risks"
- "There should be two tiers of risk thresholds, one for unacceptable and one for broadly acceptable risks."
- "There should be more than two tiers of risk thresholds."
- "There should be different risk thresholds for different types of harms (e.g. number of fatalities and economic damage)."
- "There should be multiple tiers of capabilities thresholds."
- "There should be different thresholds for different types of capabilities."
- "There should be different red lines for different types of capabilities. (Note that, by definition, there can only be one tier of red lines.)"
- "There should be a single training compute threshold"
- "There should be at least two tiers of training compute thresholds."
- "There should be separate thresholds for open models, i.e. models with broadly available model weights."

8. What are potential challenges of setting and evaluating thresholds?

- "The process of setting thresholds can be influenced by conflicts of interest."
- "Thresholds can be subject to Goodhart's Law (when a measure becomes a target, it ceases to be a good measure)"
- "AI companies may (intentionally or unintentionally) downplay the risks from their AI systems to stay below risk thresholds."
- "The lack of reliable risk estimation methods makes it challenging to set and evaluate risk thresholds."
- "Due to the unique challenges of advanced AI systems, existing risk thresholds from other safety-critical industries might not be applicable to an AI context."
- "The lack of agreed upon threat models for many risks and the variety of threat models for general purpose systems make it challenging to set capabilities thresholds."
- "The limitations of existing model evaluations make it challenging to set and evaluate capabilities thresholds."

9. When should it be evaluated whether AI systems exceed any thresholds?

- "It should be evaluated before training whether AI systems exceed any thresholds."
- "It should be evaluated during training whether AI systems exceed any thresholds."
- "It should be evaluated before deployment whether AI systems exceed any thresholds."
- "It should be evaluated after deployment whether AI systems exceed any thresholds."
- "It should be evaluated in regular intervals (e.g. every X months or an increase in training compute by Y) whether AI systems exceed any thresholds (especially if systems can learn from interactions with the world)."
- "It should be evaluated whether their AI systems exceed any risk thresholds anytime a new version is released."

10. Who should verify whether AI systems exceed any thresholds?

- "AI companies should self-assess whether their AI systems exceed any thresholds."
- "State actors should verify whether AI systems exceed any thresholds."
- "Unaccredited or accredited third-party auditors should verify whether AI systems exceed any thresholds."
- "Accredited third parties should verify whether AI systems exceed any thresholds."

11. What actions may be warranted if thresholds are exceeded?

- "If any thresholds are exceeded, AI companies should make it public."
- "If risk thresholds are exceeded, AI companies should further scrutinize the AI system before deploying it (e.g. via model evaluations)."
- "Due to the unreliability of current risk estimation methods, exceeding risk thresholds should not be part of a strict decision rule (e.g. if an AI system exceeds a risk threshold, it may not be deployed)."
- "If capabilities thresholds are exceeded, AI companies should implement additional mitigations."
- "If capabilities thresholds are exceeded and AI companies cannot implement adequate mitigations, they should pause the development and deployment process until the thresholds are no longer exceeded."
- "If capabilities thresholds are exceeded, AI companies should prepare a safety case, i.e. a report that makes a structured argument, supported by evidence, that a system is sufficiently safe."
- "If capabilities thresholds are exceeded, AI companies should notify their board of directors."
- "If capabilities thresholds are exceeded, AI companies should notify the government."
- "If capabilities thresholds are exceeded, AI companies should notify an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute)."
- "If red lines are crossed, AI companies should pause the development and deployment process until the red lines are no longer crossed."
- "If red lines are crossed, AI companies should notify their board of directors."
- "If red lines are exceeded, AI companies should notify the government."
- "If red lines are exceeded, AI companies should notify an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute)."
- "If training compute thresholds are exceeded, AI companies should conduct additional risk assessments (e.g. via model capabilities)."
- "If training compute thresholds are exceeded, AI companies should notify the government."
- "If training compute thresholds are exceeded, AI companies should notify an independent public body (e.g. the EU AI Office, FTC, or an AI Safety Institute)."

12. What mitigations would be adequate?

- "It is still unclear what specific mitigations would be adequate for specific levels of capabilities."
- "The responsibility to identify and implement adequate mitigations should be assigned primarily to AI companies."
- "AI companies should be incentivised to refine existing and develop new mitigations."

13. How should thresholds change as AI systems become more capable?

- "As AI systems become more capable, it should be assessed more rigorously whether thresholds have been exceeded."
- "As AI systems become more capable, verification procedure should be standardised such that trained auditors with sufficient access can verify whether thresholds have been exceeded."
- "As AI systems become more capable, risk thresholds should not change."
- "As AI systems become more capable, capabilities thresholds should rely more on model propensities, i.e. not just on what a system can do, but also its inclination to do these things."
- "As AI systems become more capable, actors who set capabilities thresholds should become more risk-averse, i.e. they should be more willing to accept false positives (overestimating system risks) and less willing to accept false negatives (underestimating system risks)."