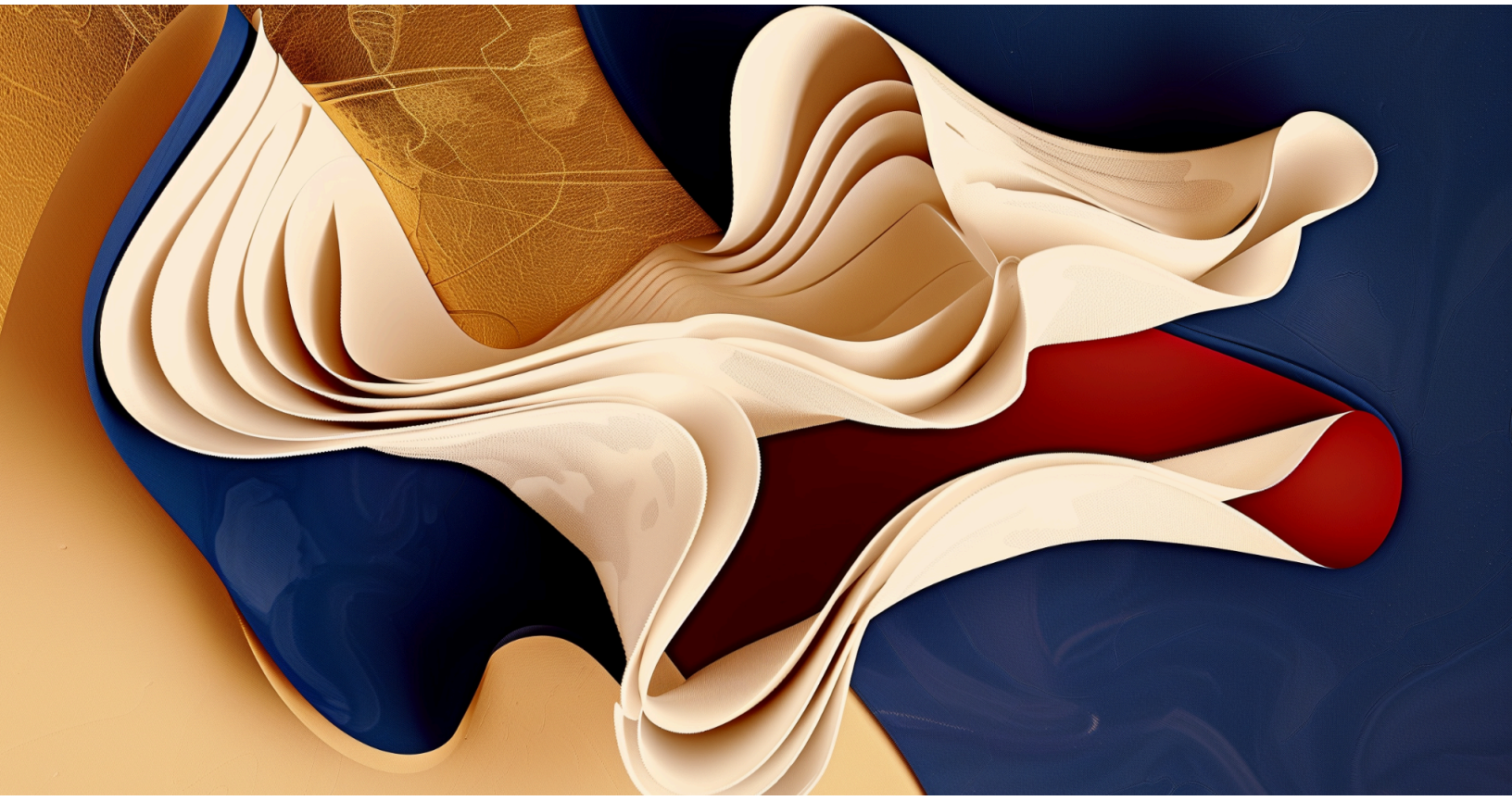
POLICY BRIEF | OCTOBER 2025

Assessing Risk Relative to Competitors: An Analysis of Current AI Company Policies

Sophie Williams, Noemi Dreksler, Aidan Homewood, Markus Anderljung, Jonas Freund



GovAI

Executive summary

- When frontier AI companies assess the risks from their models, they increasingly focus on marginal risk. This aims to measure how much their models increase risk compared to some baseline (e.g. a situation where an actor doesn't have access to any AI model). In general, this is a sensible approach.
- Recently, some companies have started to assess marginal risk relative to their competitors' models. These companies have provisions in their safety frameworks that allow them to lower their mitigations if a competitor has released a model with similar capabilities but weaker mitigations. The basic idea is that this wouldn't meaningfully increase the total level of risk in the ecosystem.
- We discuss ways in which this approach could, in fact, increase the total level of risk. For example, in practice, it's very difficult to accurately evaluate the capabilities of a competitor's model or to be confident that its mitigations are indeed weaker. As a result, companies may wrongly conclude that they can lower their own mitigations without increasing risk.
- Furthermore, the total level of risk would only stay the same under certain conditions. For example, accident risks – such as models giving adverse health advice or assisting suicide attempts – are likely to scale with the amount of users. As such, if a company were to lower their mitigations in response to others doing so it's likely to increase risk. We also identify reasons why other types of risk could increase.
- Even if these conditions hold, assessing marginal risk relative to one's competitors could still erode safety standards. Such erosion might happen suddenly, if a single defector triggers many other companies to lower their mitigations at the same time, or through a "boiling frog" effect, where small, incremental increases in risk go unnoticed, but accumulate over time.
- Given these concerns, this particular approach to marginal risk needs further scrutiny and should not be accepted as a best practice without significant analysis and public discussion.

Introduction

Frontier AI companies often assess the marginal risk of their models. This means they look at how much risk their models pose relative to some baseline (e.g. a situation where an actor doesn't have access to any AI model, or they only have access to models that were available in 2023). This is often appropriate because it tells us how much a model increases risk.

Several companies – Anthropic, OpenAI, Google DeepMind, Meta, Microsoft, and Amazon – indicate in their safety frameworks that they consider marginal risk relative to their competitors' models. In practice, this could mean that if a competitor releases a model with similar capabilities but weaker mitigations, they may decide to lower their own mitigations.

Anthropic and OpenAI justify this approach by saying that it allows them to reduce mitigations in cases where stricter measures wouldn't meaningfully reduce the total level of risk in the ecosystem.[1] To make this claim credibly, they would need to assess the other model's capabilities and verify that its mitigations are indeed weaker. In practice, companies may lack the information needed to do this.

Even if companies could reliably assess other models, the total level of risk in the ecosystem would only stay the same under certain conditions. One condition is that risk doesn't increase with the number of users exposed to "risky" models – which seems unlikely. For example, the risk of adverse mental health impacts (e.g. "AI psychosis") is likely to increase with the number of users. Another condition, which might also fail to hold, is that risks don't compound. For instance, two models with slightly different capability profiles could be used together in more harmful ways than any single model alone (Jones et al., 2024).

Even if these conditions hold, this approach could still erode safety standards. This is because companies may deem a small increase in marginal risk to be acceptable, but these increases could accumulate over time. This erosion is even more likely if companies are failing to assess marginal risk accurately. Therefore, this practice warrants closer scrutiny to ensure it's used responsibly.

Series overview. This brief is the first in a series. It examines why assessing marginal risk relative to one's competitors might be problematic. Future briefs will explore how industry could mitigate these issues, and consider potential policy responses where voluntary efforts fall short.

This first brief addresses five questions:

- What is marginal risk?
- Which companies assess marginal risk relative to their competitors?
- How do companies justify this approach?
- Why might their justification be flawed?
- Why is this a problem?

[1] Other companies don't provide a detailed justification for their approach.

What is marginal risk?

Marginal risk refers to a difference in risk relative to a baseline. This contrasts with absolute risk which refers to the total amount of risk. In general, marginal risk is easier to measure than absolute risk.[2] Figure 1 illustrates the difference between absolute and marginal risk.

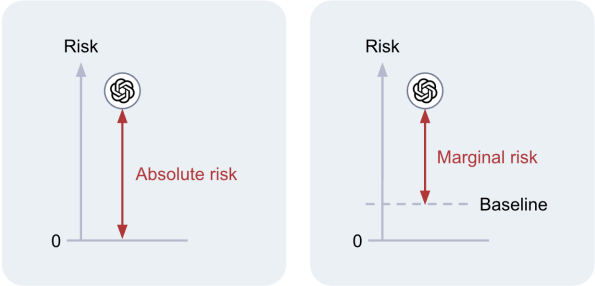


Figure 1 | Difference between absolute and marginal risk.

There are many different types of baseline. The appropriate baseline may differ depending on the context or objective. Figure 2 illustrates several possible baselines for assessing marginal risk in frontier AI. These include:

- A "human" baseline – where a model is evaluated relative to human performance on specific tasks (Wei et al., 2025).
- The state of the world at a previous point in time – for example, where a current model is evaluated relative to models available in 2023 (Alaga & Chen, 2025; Frontier Model Forum, 2025).
- The company's own previous model – where a newly released model is assessed relative to an earlier version (AISI, 2024).
- A competitor's model – which is the baseline this brief focusses on.

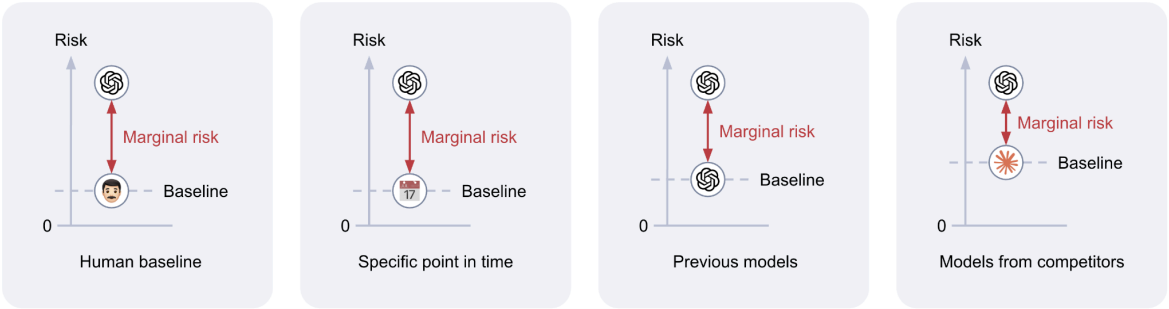


Figure 2 | Example baselines for marginal risk in the context of frontier AI.

[2] The absolute risk associated with a model is difficult to quantify because catastrophic outcomes are rare, context-dependent, and insufficiently grounded in empirical evidence. In contrast, relative risk can allow for meaningful comparisons to be made, even when absolute probabilities are uncertain.

It's unusual to measure marginal risk relative to one's competitors. Marginal risk is already used in other industries, such as nuclear energy and transport.[3] More recently, it's gained traction in the frontier AI community (Kapoor & Bommasani et al., 2024; Bengio et al., 2025; Alaga & Chen, 2025), with some companies indicating that they use competitors' models as a baseline. This contrasts with how marginal risk is typically used in other industries.[4] For example, it's hard to imagine an aviation company convincing a regulator that it should be able to lower its safety measures because a competitor has caused more fatalities.

Frontier AI companies' approach to marginal risk warrants closer scrutiny. The Frontier Model Forum highlighted that companies' use of marginal risk could lead to "risk creep" and suggested that harmonised industry practices may be beneficial (Frontier Model Forum, 2025). However, what such practices should entail remains unclear. The EU AI Act also leaves ambiguity over whether companies may assess marginal risk relative to their competitors. While this framing appears to be permissible for security risks when assessed relative to open models – as set out in Commitment 6 of the EU GPAI Code of Practice – it's less clear whether this is also true for safety risks.[5]

Which companies assess marginal risk relative to their competitors?

As of October 2025, thirteen organisations have published a safety framework.[6] At least six of these explicitly or implicitly refer to the concept of marginal risk relative to one's competitors. These are summarised below, with the relevant excerpts in the Appendix.

- **Anthropic.** Anthropic's original Responsible Scaling Policy made only a narrow reference to marginal risk defined relative to competitors' models (Anthropic, 2023). It specified that in an "extreme emergency" – for example, if "a clearly bad actor is scaling in so reckless a manner that it is likely to lead to imminent global catastrophe if not stopped (and where AI itself is helpful in such defense)" – then the company might loosen its own restrictions. Even then, it stressed that "such action would only be taken in consultation with governmental authorities". This can be seen as a precursor to the broader reference introduced in its October 2024 update (Anthropic, 2024), which remained in its most recent May 2025 revision (Anthropic, 2025a). The updated policy states that if another actor "will pass, or be on track to imminently pass" one of Anthropic's thresholds without implementing

[3] A prominent concept in European transport regulation is GAMAB ("Globalement Au Moins Aussi Bon" in French), which translates to "globally at least as good." This principle "dictates that a newly introduced system must not be more dangerous than the existing state of the art" (Tchiehe & Gauthier, 2017).
[4] This reasoning is occasionally seen in fields such as biosecurity research. If multiple labs already work with a virus, the added risk of another doing so may be considered relatively low. In contrast, work with a novel pathogen would introduce new risks that are more likely to be deemed unacceptable. However, biosecurity research operates under stringent regulatory oversight unlike frontier AI.
[5] The Code of Practice distinguishes between "safety mitigations" (Commitment 5) and "security mitigations" (Commitment 6). We use "safety risks" and "security risks" to describe the respective risks that these mitigations are intended to address.
[6] This includes Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, Nvidia, and Shanghai AI Lab.

"equivalent" mitigations, then it might lower its own. In such cases, it would advocate for US regulatory intervention to ensure the total level of risk is reduced to an acceptable level.

- **OpenAI.** OpenAI's first Preparedness Framework didn't include the concept of marginal risk relative to competitors' models (OpenAI, 2023). The concept was introduced in its second and most recent version published in April 2025, which includes a section titled "marginal risk" (OpenAI, 2025a). OpenAI's framework states that if another model has "high or critical capability" without "comparable" mitigations, then it could "adjust accordingly" the level of mitigations it requires for its own models if doing so "does not meaningfully increase the overall risk of severe harm". It also says it would publicly acknowledge any such adjustment.

- **Google DeepMind.** Google DeepMind's initial Frontier Safety Framework didn't reference marginal risk relative to competitors' models (Google DeepMind, 2024). The concept was introduced in its second version, published in February 2025 (Google DeepMind, 2025a). In it, Google DeepMind said its "adoption of the protocols [...] may depend on whether such organizations across the field adopt similar protocols". This line was removed in its most recent September 2025 update (Google DeepMind, 2025b), though this version still makes some reference to the concept. For example, it states that security levels may be adjusted if "a model does not possess capabilities meaningfully different from other publicly available models that have weaker security".

- **Meta.** In February 2025, Meta published its first Frontier AI Framework (Meta, 2025). Marginal risk appears to be a core component of Meta's risk management approach because it sets thresholds based on the extent to which a model would "uniquely enable" the execution of a "threat scenario". Meta's framework goes on to say that a model only meets its "critical" threshold if the threat scenario wouldn't occur without "this particular model". However, it's unclear how exactly they operationalise this.[7]

- **Microsoft.** In February 2025, Microsoft published its first Frontier Governance Framework (Microsoft, 2025). The framework notes that Microsoft considers "the marginal capability uplift" that a model provides relative to other available tools and information, including "currently available open-weights models". This suggests that comparisons to competitors' models could play a role in Microsoft's assessments.

- **Amazon.** In February 2025, Amazon published its first Frontier Model Safety Framework (Amazon, 2025). Although Amazon's framework says little about marginal risk relative to other models, it does say that it assesses whether a model provides a "material 'uplift' in excess of other publicly available research or existing tools", which might include other models.

[7] Meta also states that it considers the potential benefits – not just the risks – when deciding how to release models that fall below its critical threshold. This ties into the marginal risk dynamic because, by considering the potential benefits, Meta may be more inclined to conclude that a model can be deployed with lower mitigations. Given other companies assess marginal risk relative to other models (including Meta's), this could trigger them to lower their own mitigations.

How do companies justify this approach?

Anthropic and OpenAI provide some justification for this particular approach to marginal risk in their safety frameworks, whereas other companies don't explicitly do so.[8]

- **Anthropic.** Anthropic's current safety framework justifies the approach by saying that in a scenario where another actor has passed a threshold and not implemented equivalent mitigations, then the incremental increase in risk would be "small" if it were to lower its own mitigations (Anthropic, 2025a).

- **OpenAI.** OpenAI's current safety framework says that if another actor has passed a threshold without equivalent mitigations, then this would "limit the degree" to which it could reduce risk (OpenAI, 2025a). OpenAI also says it would only lower its mitigations if it could assess that this would "not meaningfully increase the overall risk of severe harm". In addition, it says it would keep its mitigations "at a level more protective than the other AI developer" and share information to validate this claim.

Intuitively, it seems reasonable for a company to scale back mitigations when a similarly capable model lacks equivalent measures. In such cases, maintaining higher mitigations on one's own model may not meaningfully reduce the total level of risk in the ecosystem. From this perspective, a company might ask: why invest additional time and resources into mitigations that don't make the world appreciably safer?

Companies may also be concerned that stricter mitigations put them at a competitive disadvantage. Although this concern isn't explicitly stated in any of the safety frameworks, mitigations can be costly and time-consuming.[9] Benchmarking against one's competitors therefore allows companies to reduce costs when others aren't holding themselves to the same standard. This consideration is especially salient in the context of a "high-stakes, global technology race" (Karnofsky, 2024).

Why might their justification be flawed?

As discussed in the last section, Anthropic and OpenAI suggest that this approach to marginal risk allows them to lower mitigations in cases where doing so wouldn't meaningfully increase the total level of risk in the ecosystem.

Verifying this claim is very difficult. The process can be broken down into three high-level steps (see Figure 3), each of which present significant challenges. This creates a problematic dynamic

[8] Other companies may justify or operationalise the concept differently, but their safety frameworks don't provide enough detail for us to analyse their approach in the same level of depth.
[9] Despite this concern, it's important to note that higher safety standards can also confer competitive advantages. First, companies known for safety leadership may earn a "trust premium", making it easier to attract customers, lucrative partnerships, and top talent. Second, proactive safety measures can position companies ahead of future regulatory requirements, giving them greater influence over how regulations are designed and implemented. Third, a strong safety record can help protect against reputational damage in the event of an incident.

where critical safety decisions might be based on inaccurate assessments. Notably, companies' safety frameworks don't say how they carry out these steps nor how they deal with the challenges involved.

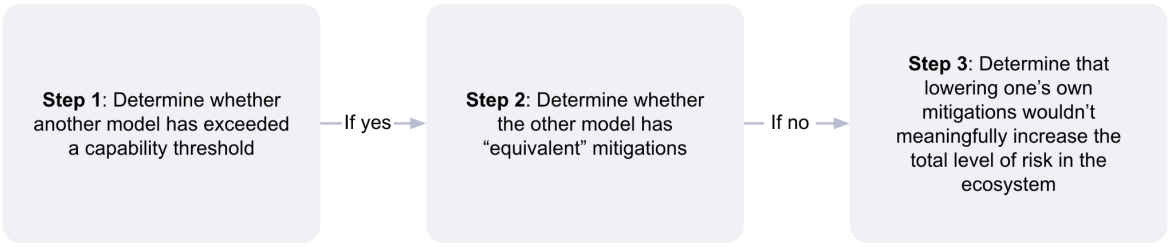


Figure 3 | High-level steps in assessing marginal risk relative to competitors' models.

Step 1: Determine whether another model has exceeded a capability threshold. If done inaccurately, there's a risk of false positives – where a company overestimates another model's capabilities, and uses this to justify lowering its own mitigations. We may see such false positives because:

- Even assessing the capabilities of one's own model is difficult. This was illustrated by Anthropic's decision to deploy Claude Opus 4 with ASL-3 measures (Anthropic, 2025b). In its press release, Anthropic acknowledged that it was still determining whether Claude Opus 4 had definitively passed the threshold requiring ASL-3 protections. It said this uncertainty arose because capability evaluations are "inherently challenging" and that, as models approach higher thresholds, it "takes longer to determine their status".
- If a company relies on what other companies have said publicly about their models – for example, in model cards, research papers, or press releases – this may present an incomplete picture, rely on non-comparable evaluation methods or benchmarks, or even exaggerate model performance.[10]

Step 2: Determine whether the other model has equivalent mitigations.[11] If done inaccurately, there's a risk of false negatives – where a company underestimates the effectiveness of their competitors' mitigations, and uses this to justify lowering its own mitigations. We may see such false negatives because:

- Companies may not share information about certain mitigations for competitive, security, or confidentiality reasons.
- Observing a model's behaviour may not reveal the full range of mitigations in place (e.g. any pre-deployment measures).

[10] Although note that there have been instances of companies collaborating on evaluations which could support such assessments (see, e.g. OpenAI, 2025b).
[11] Note that the terminology used by companies in their safety frameworks varies (e.g. Anthropic uses the term "equivalent", whereas OpenAI uses the term "comparable").

- Techniques such as constitutional AI, filtering, and reinforcement learning from human feedback (RLHF) may vary in implementation, but still achieve equivalent outcomes.
- Companies appear to lack rigorous frameworks for evaluating the effectiveness of their own mitigations.[12] Without such benchmarks, it's unlikely they could reliably assess others' mitigations or determine their equivalence.

Step 3: Determine that lowering one's own mitigations wouldn't meaningfully increase the total level of risk in the ecosystem. To do this, companies would need to establish what counts as a "meaningful" increase in total risk.[13] They would also need to identify which mitigations could be reduced – and by how much – without increasing the total level of risk beyond these bounds. Whether this is possible depends on whether certain conditions hold. This varies depending on the type of risk in question. For example:

- **Misuse risks.** Consider a potential misuse scenario in which a terrorist wants to use a model to help develop a biological weapon. A company might argue that if such a model is already available, then lowering one's own mitigations for biological misuse won't meaningfully increase risk. However, this reasoning depends on conditions that may not hold in practice (see Table 1). For example, bad actors might not know which model is "riskiest" and instead use a model that's more well-known, but less capable. If the mitigations on the more well-known model are lowered, the bad actor may be more likely to succeed in developing a weapon. The same logic applies to other misuse risks, such as cyberattacks.

- **Accident and structural risks.** For accident risks – harms from models behaving in unintended ways – total risk would only stay the same if reducing one's mitigations doesn't increase overall exposure. In other words, the risk of AI models inducing negative mental health effects (e.g. "AI psychosis"), assisting suicide attempts, or giving adverse health advice could increase if another company decided to lowers its mitigations because the total number of exposed users would be higher. Structural risks – arising from the way models reshape the systems, incentives, and environments in which they are deployed – might also scale with additional deployments of "risky" models.

- **Autonomy risks.** In the case of autonomy risks – such as loss of control or misalignment – it seems even less likely that the total level of risk would be meaningfully unchanged, because the potential for compounding effects is high. For example, if one "risky" model carries a 1% chance of a loss-of-control event, then two such models could double that probability. Worse, if several "risky" models were to escape, they might interact in unpredictable ways. For example, they could inadvertently amplify each other's harmful behaviours or intentionally exchange code and strategies, creating hybrid systems that are even more misaligned and difficult to control.

[12] Approaches have been proposed (AISI, 2025a, AISI, 2025b), but it's unclear whether these are common practice.
[13] Note that the terminology used by companies in their safety frameworks varies (e.g. OpenAI uses the term "meaningful", whereas Anthropic uses the term "small").

Condition	Description	Validity
Awareness	Bad actors are aware of the competitor's "risky" model	🟡 **Questionable** – Bad actors might not realise that there's already a model that's able to help them with a malicious task. For example, information about the models' dangerous capabilities might not be publicly available. Therefore, lowering one's own mitigations might increase the chance that bad actors become aware of a model that's more capable of helping them.
Low switching costs	Bad actors are able to switch to the competitor's "risky" model with little friction	🟡 **Questionable** – Switching costs are probably low in most cases, but not always. Models can vary in terms of how easy they are to use (e.g. due to computational requirements or integration complexity). In particular, switching from a "closed" to an "open" model might pose relatively high costs. Bad actors might find it difficult to switch to the competitor's model for these reasons, but find it easier to switch to yours (or already be using it).
No complementarities	Using two "risky" models together provides no additional benefit to bad actors	🟠 **Unlikely** – Different models may have complementary strengths that, when used together, enable more harmful outcomes than any single model alone (Jones et al., 2024). Therefore, lowering one's own mitigations might provide additional benefits to the bad actor if they use the two models alongside each other.

Table 1 | Conditions that would need to hold for misuse risk to stay the same. Conditions required for the total level of misuse risk to stay the same if a company's own mitigations were lowered to those of a competitor, provided the models have the same level of capabilities.[14]

Bias can also increase the risk of error. Cognitive and institutional biases can complicate this assessment even further. Biases may cause companies to overestimate the effectiveness of their own safety measures or underestimate those of their competitors. Competitive pressures could reinforce these tendencies, amplifying the risk of misjudgement.

[14] This list is not intended to be exhaustive.

Why is this a problem?

This approach to marginal risk raises some significant problems.

First, inaccurate assessments could lead to decisions that *do* meaningfully increase the total level of risk in the ecosystem. As discussed in the last section, it's very difficult to assess marginal risk accurately, not least because of the lack of information available about competitors' models. This could mean that decisions to lower one's own mitigations does actually result in a meaningful increase in risk. This undermines the very rationale given by Anthropic and OpenAI for taking this approach in the first place.

Second, even if companies could conduct these assessments perfectly, the approach could still erode safety standards. This erosion could manifest in a couple of ways. For example:

- **A "race to the bottom" dynamic.** When marginal risk is defined relative to one's competitors, it effectively enables the least responsible actor to set the industry baseline. A single "defector" from safety norms could therefore trigger a rapid downward shift as others lower their own mitigations to avoid being placed at a competitive disadvantage. Given intense competition in the frontier AI market, this scenario seems possible unless a government or regulatory body were able to intervene.

- **A "boiling frog" dynamic.** If multiple companies each take individual decisions to lower mitigations that only increase risk marginally, the total level of risk could still accumulate over time. Each company may reasonably judge that its own contribution to total risk is small, but the cumulative effect when repeated across the industry could be substantial. Because each decision appears justified in isolation, the gradual increase in total risk may go unnoticed – much like a frog failing to sense the water slowly heating up. This scenario also seems plausible, especially without a regulator or other body monitoring the total level of risk in the ecosystem.

Beyond risk-related considerations, there are other reasons why this approach may be undesirable. Specifically, companies may be concerned that:

- **It could weaken public trust**. If companies lower their mitigations, public confidence in responsible frontier AI development may decline – potentially triggering a backlash against the industry. The likelihood of this depends on how the dynamic unfolds – whether as a dramatic "race to the bottom", a slow "boiling frog" effect, or something in between. If risk increases gradually, a public backlash may be less immediate, but the trend could still attract attention from researchers, advocacy groups, or the media, particularly if it contributes to a visible incident.

- **It could increase legal liability**. The case for liability may be particularly strong if a company initially recognises a risk and implements mitigations, but later weakens them in response to a competitor's behaviour. In such circumstances, the prior acknowledgement of risk

undermines any argument that the risk was “unforeseeable”. Liability could also arise where harm is caused to a single party and multiple models contribute to that harm (e.g. in the generation of non-consensual intimate imagery). Consider the following analogy: a worker works at five different asbestos mines over their career. Exposure at any two mines is sufficient to cause mesothelioma. Each company could argue that because the others also failed to provide protection, their own conduct didn’t increase the marginal risk – the workers would develop mesothelioma regardless. Yet, in some jurisdictions, courts would hold each company liable for failing to protect their workers, even though precautions by any single company wouldn’t have prevented the harm. Some tort law regimes recognise this through doctrines that abandon “but-for” causation tests in cases involving multiple contributors to harm. In principle, this appears to be a real risk for companies, though a full legal analysis is beyond the scope of this brief.

Conclusion

While the concept of marginal risk has valid applications in frontier AI risk management, defining it relative to one’s competitors raises some important concerns. Most notably, it could erode safety standards and increase the total level of risk in the ecosystem. The approach is also underdeveloped in companies’ safety frameworks and underexplored in the academic literature. Given these shortcomings, further work is needed to establish how this particular approach to marginal risk can be used responsibly. Companies, industry bodies, and policymakers are all likely to have roles to play in this process, which we will explore in subsequent policy briefs.

About the Authors

Sophie Williams

Research Fellow, Centre for the Governance of AI

Sophie works across GovAI's risk management and policy workstreams. Before joining GovAI, she advised government ministers at the UK Home Office and the Department for Science, Innovation & Technology. She has also held roles at the Financial Conduct Authority and the Digital Regulation Cooperation Forum. Sophie holds a BA in Experimental Psychology from the University of Oxford.

Noemi Dreksler

Senior Research Fellow, Centre for the Governance of AI

Noemi is a member of GovAI's Risk Management Team. Before joining the team, Noemi's work at GovAI focused on exploring public, expert, and elite opinion on AI. She holds a DPhil in Experimental Psychology from the University of Oxford, an MSc in Organisational Psychology from UCL, and a BA in Psychology and Philosophy from the University of Oxford.

Aidan Homewood

Research Scholar, Centre for the Governance of AI

Aidan researches risk management for frontier AI. He is particularly interested in safety frameworks and external assurance. Before joining GovAI he advised New Zealand policymakers on AI policy, was a Fellow at Pivotal Research, and co-founded an edtech startup. He holds a BSc in Mathematics from Victoria University of Wellington.

Markus Anderljung

Director of Policy and Research, Centre for the Governance of AI

Markus leads research at GovAI. He served as one of the Vice-Chairs drafting the EU's Code of Practice for General Purpose AI, and was previously seconded to the UK Cabinet Office as a Senior AI Policy Specialist. He is also an Adjunct Fellow at the Center for a New American Security and a member of the OECD AI Policy Observatory's Expert Group on AI Futures.

Jonas Freund

Senior Research Fellow, Centre for the Governance of AI

Jonas leads GovAI's Risk Management Team. Before joining GovAI, he advised the UK Government on AI regulation, interned at Google DeepMind's Public Policy team, and helped found the Institute for Law and AI (LawAI), where he is still a board member. He holds a law degree from Heidelberg University and a PhD in law from Goethe University Frankfurt.

Acknowledgements

We are grateful for valuable comments and feedback from Adam Bales, Alan Chan, Anne le Roux, Ben Clifford, Ben Garfinkel, Jack Wadham, Jide Alaga, John Halstead, Marie Buhl, Matthew van der Merwe, Michael Chen, Nicola Ding, Tammy Masterson, Tommy Shaffer Shane, and Zaheed Kara. All remaining errors are our own.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in London, UK. It was founded in 2018 at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to help decision-makers navigate the transition to a world with advanced AI, by producing rigorous research and fostering talent. The central focus of our research is threats that general-purpose AI systems may pose to security. We seek to understand the risks they pose today, while also looking ahead to the more extreme risks they could pose in the future.

Appendix: Excerpts from frontier AI safety frameworks

Anthropic

> *"**It is possible** [...] **that another actor in the frontier AI ecosystem will pass, or be on track to imminently pass, a Capability Threshold without implementing measures equivalent to the Required Safeguards such that their actions pose a serious risk for the world**. In such a scenario, **because the incremental increase in risk attributable to us would be small, we might decide to lower the Required Safeguards**. If we take this measure, however, we will also acknowledge the overall level of risk posed by AI systems (including ours), and will invest significantly in making a case to the U.S. government for taking regulatory action to mitigate such risk to acceptable levels."* (Anthropic, 2025a, p. 13)

OpenAI

> *"We recognize that **another frontier AI model developer might develop or release a system with High or Critical capability in one of this Framework's Tracked Categories and may do so without instituting comparable safeguards to the ones we have committed to**. **Such an action could significantly increase the baseline risk of severe harm being realized in the world, and limit the degree to which we can reduce risk using our safeguards**. If we are able to rigorously confirm that such a scenario has occurred, then **we could adjust accordingly the level of safeguards that we require in that capability area**, but only if: we assess that doing so does not meaningfully increase the overall risk of severe harm, we publicly acknowledge that we are making the adjustment, and, in order to avoid a race to the bottom on safety, we keep our safeguards at a level more protective than the other AI developer, and share information to validate this claim."* (OpenAI, 2025a, p. 12)

Google DeepMind

> *"**Importantly, there are certain mitigations whose social value is significantly reduced if not broadly applied to frontier AI models reaching critical capabilities.** These mitigations are most effective when adopted by industry as a whole: our adoption of them would result in effective risk mitigation for society only if all relevant organisations provide similar levels of protection."* (Google DeepMind, 2025b, p. 2)

> *"**These recommended security levels reflect our current thinking proportionate to the risks posed and may be adjusted if our understanding of the risks changes. This may occur if, for example, a model does not possess capabilities meaningfully different from other publicly available models that have weaker security applied** (in which case the marginal benefit of higher security is limited), or if we assess that the benefits of the open release of model weights outweigh the risks. Relatedly, **we believe these recommendations will only be effective if the entire frontier AI field applies them, and of limited social utility if not**."* (Google DeepMind, 2025b, p. 9)

Meta

> *"We define our thresholds based on the extent to which frontier AI would **uniquely enable** the execution of any of the threat scenarios we have identified as being potentially sufficient to produce a catastrophic outcome."* (Meta, 2025, p. 4)
>
> *"The term "uniquely enabling" is defined in the appendix: "Uniquely enabling describes a model that is an essential controlling factor in a given outcome. **A model is considered to meet the critical risk threshold if it is determined that a specified threat scenario would not occur without this particular model**."* (Meta, 2025, p. 20)

Microsoft

> *"Holistic risk assessment: The results of capability evaluation and an assessment of risk factors external to the model then inform a determination as to whether a model has a tracked capability and to what level. This includes assessing the impact of potential system-level mitigations and societal and institutional factors that can impact whether and how a hazard materializes. **This holistic risk assessment also considers the marginal capability uplift a model may provide over and above currently available tools and information**, including currently available open-weights models."* (Microsoft, 2025, p. 6)

Amazon

> *"The CBRN Capability Threshold focuses on the potential that a frontier model may provide actors **material "uplift" in excess of other publicly available research or existing tool**s, such as internet search."* (Amazon, 2025, p. 2)
>
> *"The Offensive Cyber Operations Threshold focuses on the potential that a frontier model may provide **material uplift in excess of other publicly available research or existing tools**, such as internet search."* (Amazon, 2025, p. 2)

References

AI Security Institute. (2024, November 2024). *Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet*. https://www.aisi.gov.uk/blog/pre-deployment-evaluation-of-anthropics-upgraded-claude-3-5-sonnet

AI Security Institute. (2025a, February, 4). *Principles for Safeguard Evaluation*. https://www.aisi.gov.uk/blog/principles-for-safeguard-evaluation

AI Security Institute. (2025b, May 29). *Making Safeguard Evaluations Actionable*. https://www.aisi.gov.uk/blog/making-safeguard-evaluations-actionable

Amazon. (2025, February 10). *Amazon's Frontier Model Safety Framework*. https://assets.amazon.science/a7/7c/8bdade5c4eda9168f3dee6434fff/pc-amazon-frontier-model-safety-framework-2-7-final-2-9.pdf

Anthropic. (2023, September 19). *Anthropic's Responsible Scaling Policy* (Version 1.0). https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

Anthropic. (2024, October 15). *Responsible Scaling Policy* (Version 2.0). https://www-cdn.anthropic.com/616dee633636e5bd309cb73aed8622e80fe47839.pdf

Anthropic. (2025a, May 14). *Responsible Scaling Policy* (Version 2.2). https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

Anthropic (2025b, May 22). *Activating AI Safety Level 3 protections*. https://www.anthropic.com/news/activating-asl3-protections

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... Dietterich, T. G. (2025). *International AI Safety Report*. arXiv. https://arxiv.org/abs/2501.17805

Chen, M. & Alaga, J. (2025). *Marginal Risk Relative to What? Distinguishing Baselines in AI Risk Management*. ICML Workshop on Technical AI Governance. https://openreview.net/pdf/97af56981533ce22d1e59af3085e73b8ccfe345a.pdf

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

Frontier Model Forum. (2025, June 18). *Risk Taxonomy and Thresholds for Frontier AI Frameworks*. https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds

Google. (2024, May 17). *Frontier Safety Framework* (Version 1.0). https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf

Google. (2025a, February 4). *Frontier Safety Framework* (Version 2.0). https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0.pdf

Google. (2025b, September 22). *Frontier Safety Framework* (Version 3.0). https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf

Jones, E., Dragan, A., Stainhardt, J. (2024). *Adversaries can Misuse Combinations of Safe Models*. arXiv https://arxiv.org/abs/2406.14595

Kapoor, S. & Bommasani, R. et al., (2024). *On The Societal Impact of Open Foundation Models*. arXiv. https://arxiv.org/abs/2403.07918

Karnofsky, H. (2024, September 13). *If-Then Commitments for AI Risk Reduction*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction

Meta. (2025, February 3). *Frontier AI Framework* (Version 1.1). https://ai.meta.com/static-resource/meta-frontier-ai-framework

Microsoft. (2025, February 8). *Frontier Governance Framework*. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

OpenAI. (2023, December 18). *Preparedness Framework* (Beta). https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenAI. (2025a, April 15). *Preparedness Framework* (Version 2). https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

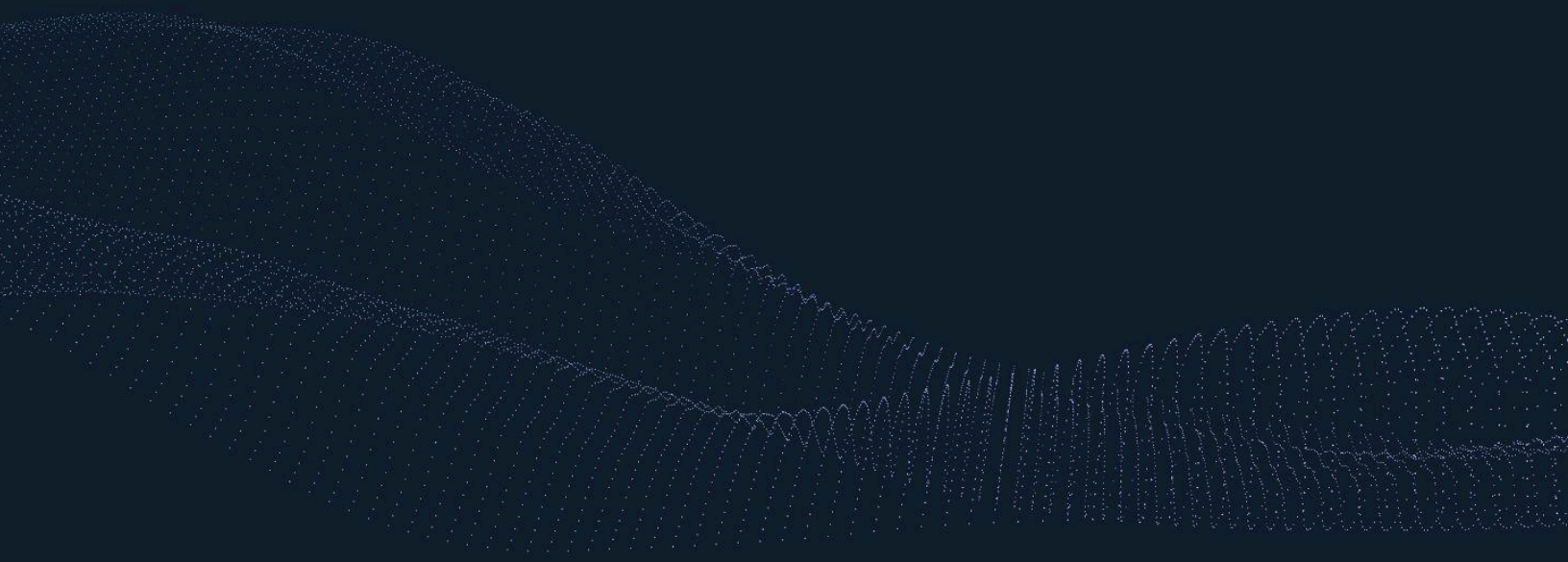
OpenAI. (2025b, August 27). *Findings from a pilot Anthropic–OpenAI alignment evaluation exercise: OpenAI Safety Tests.* https://openai.com/index/openai-anthropic-safety-evaluation/

Regulation (EU) 2024/1689. *Laying Down Harmonised Rules on Artificial Intelligence* (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj

Tchiehe, D. & Gauthier, F. (2017). Classification of Risk Acceptability and Risk Tolerability Factors in Occupational Health and Safety. *Safety Science, 92*, 138–147. https://doi.org/10.1016/j.ssci.2016.10.003

Wei, K., Paskov, P., Dev. S., Byun, M., Reuel, A., Roberts-Gaal, X., Calcott, R., Coxon, E., Deshpande, C. (2025). *Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations*. arXiv. https://arxiv.org/abs/2506.13776

Zwetsloot, R. & Dafoe, A. (2019, February 11). *Thinking About Risks From AI: Accident, Misuse and Structure*. Lawfare. https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure



GovAI