

---

# Declare and Justify: Explicit assumptions in AI evaluations are necessary for effective regulation

---

**Peter Barnett**                      **Lisa Thiergart**  
Machine Intelligence Research Institute  
Berkeley, CA, USA 94704  
{peter, lisa}@intelligence.org

## Abstract

As AI systems advance, AI evaluations are becoming an important pillar of regulations for ensuring safety. We argue that such regulation should require developers to **explicitly identify and justify key underlying assumptions about evaluations** as part of their case for safety. We identify core assumptions in AI evaluations (both for evaluating existing models and forecasting future models), such as comprehensive threat modeling, proxy task validity, and adequate capability elicitation. Many of these assumptions cannot currently be well justified. If regulation is to be based on evaluations, it should require that AI development be halted if evaluations demonstrate unacceptable danger or if these assumptions are inadequately justified. Our presented approach aims to enhance transparency in AI development, offering a practical path towards more effective governance of advanced AI systems.

## 1 Introduction

The rapid pace of AI development has prompted demands for regulation to help safeguard against novel risks, including catastrophic risks [1]. This regulation should aim to prevent harms caused by malicious actors misusing AI systems, as well as large-scale accident risks caused by autonomous AI systems acting in misaligned ways [2–4].

Today’s frontier AI systems are not created by understanding and implementing specific capabilities; they are instead iteratively shaped through a training process that encourages instrumental capabilities to emerge. Consequently, AI developers do not know what their systems will be capable of until they test them — and sometimes not even then. As OpenAI CEO Sam Altman said about predicting capabilities [5], “Until we go train that model, it’s like a fun guessing game for us.”

Major AI developers [6–8] have put forward plans for safety based on AI evaluations [9, 10] that attempt to assess a model’s capacity to facilitate dangerous activities such as hacking [11, 12], bioweapons design [13, 14], and human manipulation [15, 16]. Governments are requiring that AI developers provide them access to models for testing [17, 18]. Clearly, much depends on these evaluation efforts, especially for avoiding potentially catastrophic risks.

But safety cases based on AI evaluations rest on many underlying assumptions about the scope and limitations of testing that may not have been adequately interrogated. Previous work discusses various limitations to AI evaluations [19–22]. In this paper, we **identify key assumptions** we argue should be stated and justified by developers as part of any safety plan or regulatory effort.

## 2 Current AI evaluations workflow

AI evaluations form a large component of AI developer safety plans, such as the Anthropic Responsible Scaling Policy [6], the OpenAI Preparedness Framework [7], and the Google Deepmind Frontier Safety Framework [8]. These seek to estimate whether current models have dangerous capabilities that could lead to catastrophic harm, and to predict if future models will. In this paper, we distinguish between existing and future models because the capabilities of future models can only be inferred from those of existing models which can be directly interacted with. These capability evaluations can include “human uplift” studies, where the capability being measured is the ability to assist humans at harmful tasks [23, 24].

For evaluating **existing models**, the process is:

1. Assess threat vectors via which the AI system could cause harm.
2. Design proxy tasks which estimate the system’s ability to exploit these threat vectors.
3. Attempt to get the model to do these proxy tasks.
4. If a model can do these proxy tasks, trigger an action such as don’t release the model or don’t continue training the model without first resolving the risk.

An implicit assumption is made that if evaluators are unable to make a model perform well on the proxy tasks, then it is unlikely to have dangerous capabilities, and therefore will be safe to deploy.

Developers today acknowledge that some AI systems may not be safe to even create; for example, an inadequately secured model could be used to cause harm if stolen, or a model that has capabilities allowing it to break out of its containment could act autonomously. For forecasting and preventing risks from **future models**, it appears from the developer safety plans that the standard process is:

1. Assess threat vectors via which *future* AI systems could cause harm.
2. Determine precursor capabilities that would appear before an AI system develops the actually dangerous capabilities.
3. Design proxy tasks for these precursor capabilities.
4. Attempt to get the model to do these proxy tasks.
5. If a model ever displays these precursor abilities, stop development or deployment until sufficient precautions [25] are implemented.

This approach makes the implicit assumption that there exists enough of a time and compute [26] gap between reaching precursor capability and full required capability for evaluators to catch the precursors and stop further development.

## 3 Key assumptions in AI evaluations

For AI evaluations to provide justified confidence in a model’s lack of dangerous capabilities, several key assumptions must hold. Below, we explore a non-exhaustive listing of relevant assumptions.

### 3.1 Evaluating existing models

**1. Comprehensive Threat Modeling: Have all the relevant threat vectors been considered?** Evaluators must adequately cover the space of dangerous capabilities the AI system could have that would allow it to cause harm. This requires threat modeling which covers *all* exploitable threat vectors, including vectors evaluators didn’t consider but which the AI system might be capable of finding without detection.

Sufficient justification for this assumption may be obtainable when the goal is to prevent harm via misuse by malicious actors. But this would require evaluators (potentially working with threat assessors and domain experts) to be correctly confident that they can find all threat vectors that malicious actors would be able to find. AI developers have committed to working with domain experts as part of safety assessments [6, 27]. This could prove challenging, especially when considering well-resourced (potentially nation-state level) malicious actors.

**It may be substantially more difficult to justify this assumption when considering risks from autonomous AI systems;** for example, AI systems may notice threat vectors that humans do not.

**2. Proxy Task Validity: Are the proxy tasks a good measure of the dangerous tasks?** In order for the proxy tasks to be a good measure, it must be true that if an AI system does not succeed at a proxy task, the system cannot succeed at the more complex dangerous task. Therefore, success at the proxy must be a necessary requirement for success at the dangerous task. If there was an unanticipated alternative path to complete the dangerous task, the proxy would fail to provide a strong safety argument.

An assumption of proxy validity is potentially justifiable when considering misuse risk, although it requires the evaluators to understand the threat vectors as well as potential malicious actors do, which will likely require significant domain knowledge. It is not clear how to justify this assumption when considering risks from AI systems acting autonomously, because AI systems may be able to exploit threat vectors in inhuman ways that don't require competence at the proxy tasks.

**3. Adequate Capability Elicitation: Do evaluators adequately elicit model capabilities?** If evaluators are not able to elicit the AI system's full capabilities [28], this can result in false negatives, leading the evaluators to falsely believe a system to be incapable of a task. For example, if they did not use all possible post-training enhancements [29]. How can we assess whether the elicitation of a capability has been *adequate*? This depends on the threat model being evaluated:

**Misuse:** An assumption of adequate elicitation for the purposes of assessing potential misuse by malicious actors may be justifiable, although this may require asymmetrical provision of affordances: e.g. evaluators have access to non-safety-trained models and fine-tuning [30], while public users have only monitored API access to the model.

**Autonomy:** To assess if a model is capable of causing harm autonomously (i.e. misalignment risk), evaluators are not comparing against a human baseline, and must be able to elicit close to the "true limit" of a model's capabilities. This effort could fail if there are methods for capabilities elicitation unknown to the evaluators, or if the model strategically underperforms on evaluations [31]. It is currently unclear how an assumption of full elicitation of autonomous capabilities can be justified.

### 3.2 Forecasting future models

**1. Comprehensive Coverage of Future Threat Vectors: Have all the threat vectors relevant to the next iteration of models been considered?** This is similar to the assumption of comprehensive threat modeling when evaluating existing models, but with greater inherent uncertainty. Justifying this assumption requires either a comprehensive mapping of possible future threat vectors or a robust argument that newly created models will not be capable of taking advantage of novel, unanticipated threat vectors.

This assumption may be justifiable for misuse risks, although not at present given the nascent state of AI threat modeling. This would require that future models do not have unexpected or unconsidered capabilities that malicious actors could learn about and use. This assumption is not feasible to justify when considering risks from autonomous AI systems. Methods do not exist to predict if or when models will gain the ability to exploit threat vectors which evaluators may fail to consider.

**2. Validity of Precursor Capability Proxies: Are the proxy tasks a good measure of the precursor capabilities?** This assumption is again much the same as the assumption of proxy task validity for evaluating existing models, and faces similar difficulties.

**3. Necessity of Precursor Capabilities: Are the precursor capabilities necessary for the development of the dangerous capabilities?** If the precursor capabilities are not actually prerequisites for the dangerous capabilities, then AI developers may inadvertently create dangerous models because they did not observe the precursor capabilities.

Understanding of how capabilities arise in AI models is lacking, and there are not good methods for determining that certain capabilities will arise before others. While some capabilities may predictably arise before others (e.g. a model will likely be able to do basic programming before it is able to write complicated back-doored code), this assumption cannot, at present, be rigorously justified.

**4. Adequate Elicitation of Precursor Capabilities: Do evaluators adequately elicit model precursor capabilities?** This assumption is again much the same as the assumption of adequate capability elicitation for evaluating existing models, and faces similar difficulties.

**5. Sufficient Compute Gap between Precursor and Dangerous Capabilities: Is there a compute gap that is large enough to catch precursor capabilities before dangerous capabilities develop?** The gap must be large enough and the evaluation frequency high enough such that warning signs are caught before dangerous capabilities arise. Given the absence of methods to predict the size, or even the existence, of this gap, **this assumption cannot be robustly justified.** Precursor and dangerous capabilities may arise at the same time, such as if they result from an underlying third factor. There are also no guarantees that there will not be sharp capabilities jumps [32, 33], or that models will not rapidly transition from having few to all of the precursor capabilities over a small increase in scaling.

**6. Comprehensive Tracking of Capability Inputs: Are all factors which lead to increased capabilities being tracked?** To forecast model capabilities, evaluators need to be tracking all the relevant factors that will change between existing models and more capable future models. This goes beyond just tracking the total compute used in training and should include architectural changes, data quality, different training setups, and other algorithmic changes. It may be feasible for evaluators to track the inputs into AI capabilities, especially if AI developers who understand these factors are required to honestly report on these.

**7. Accurate Capability Forecasts: Are evaluators able to make accurate forecasts based on evaluations?** Probabilistic forecasts based on these evaluations must be accurate enough to reliably determine future model capabilities [34]. As well as tracking all relevant inputs contributing to capabilities, evaluators must also have a sufficiently accurate forecasting model for predicting how these inputs translate into future capabilities.

Evaluators might gain some confidence in their predictions by establishing a good track record, but currently these track records do not exist. There may not be many generations of AI models before they become dangerous, making it challenging to establish a track record. Factors such as novel algorithmic progress could also disrupt these forecasts.

## 4 Implications for regulation

AI regulation aimed at preventing catastrophic harm may amongst other components heavily rely on AI evaluations. However, as discussed, gaining assurance of safety using AI evaluations relies on many underlying assumptions. **We propose that regulation based on AI evaluations should require AI developers to publish a list of the assumptions being made (e.g. the assumptions listed in this paper) and justify them, and these justifications should be subject to review by third party experts.** Justifying these assumptions is essential as part of a case for safety, and if the assumptions are not justified it is not appropriate to make inferences about a model's capabilities beyond the specific tests. This proposal is intended as a practical measure to enhance transparency and assist regulators in determining whether AI development is safe.

Evaluations can provide useful information about model capabilities, and should likely be performed even if the assumptions cannot be justified. But evaluations should not be used to argue that AI systems are safe in the absence of such justifications. AI evaluations should not provide a false sense of security, and rigorously listing assumptions may help alleviate this.

We do not know exactly what AI regulation will look like, however regulation may be based on the capabilities of AI models; for example, models with certain capabilities may only be deployed with certain precautions, or AI developers may be required to argue that their systems are safe because they lack certain capabilities.

AI developers should explicitly state and justify the assumptions being made as part of an evaluations-based case for safety. These should be released for public scrutiny, as long as this itself would be safe (for example, it should not alert malicious actors to novel threat vectors). These assumptions and justifications should then be assessed by third-party experts. For example, if AI developers are required to publish safety and security protocols [35] which rely on evaluations, they

should also include a comprehensive list of the assumptions being made, and how or whether these are justified.

Regulation could mandate that AI development (and certain deployments) should not continue if:

- Evaluations reveal that a system has a sufficiently high probability of being dangerous, or that future systems will be dangerous before there are adequate security precautions. This could include triggering “red lines” or “yellow lines”—predefined specific capability thresholds beyond which AI development must stop or proceed only with extreme caution.
- The AI developer does not provide a list of assumptions and justifications, or if these justifications are judged to be inadequate. When dealing with high-risk AI systems, development should not continue if the assumptions are not judged to hold with very high probability.

## 5 Conclusion

In this paper, we have discussed the role of AI evaluations for avoiding catastrophic risks. We have identified key assumptions in safety cases based on AI evaluations, both for evaluating existing models and for forecasting future models. If AI regulation is to be based on such evaluations, it should require AI developers to comprehensively identify and justify these assumptions. This can offer regulators much needed transparency into determining whether AI development is safe. AI development should then be halted if evaluations reveal unacceptable dangers, or if the underlying assumptions cannot be adequately justified.

## References

- [1] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier AI Regulation: Managing Emerging Risks to Public Safety, November 2023. arXiv:2307.03718 [cs].
- [2] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845, May 2024. Publisher: American Association for the Advancement of Science.
- [3] Yoshua Bengio, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, et al. *International Scientific Report on the Safety of Advanced AI*. Department for Science, Innovation and Technology, 2024.
- [4] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*, 2023.
- [5] M Murgia. OpenAI chief seeks new Microsoft funds to build ‘superintelligence.’. *The Financial Times*, 2023.
- [6] Anthropic. Anthropic’s Responsible Scaling Policy, 2023.
- [7] OpenAI. Preparedness Framework (Beta), 2023.
- [8] Google Deepmind. Frontier Safety Framework, 2024.
- [9] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, September 2023. arXiv:2305.15324 [cs].

- [10] Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating Frontier Models for Dangerous Capabilities, April 2024. arXiv:2403.13793 [cs].
- [11] Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang. LLM Agents can Autonomously Exploit One-day Vulnerabilities, April 2024. arXiv:2404.08144 [cs].
- [12] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, June 2024. arXiv:2406.01637 [cs].
- [13] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. Can large language models democratize access to dual-use biotechnology?, June 2023. arXiv:2306.03809 [cs].
- [14] Jaspreet Pannu, Doni Bloomfield, Alex Zhu, Robert MacKnight, Gabe Gomes, Anita Cicero, and Thomas Inglesby. Prioritizing High-Consequence Biological Capabilities in Evaluations of Artificial Intelligence Models, May 2024.
- [15] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, March 2024. arXiv:2403.14380 [cs].
- [16] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, February 2024.
- [17] U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI — nist.gov. <https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research-testing-and-evaluation-with-anthropic-and-openai>. [Accessed 12-09-2024].
- [18] Vincent Manancourt, Gian Volpicelli, and Mohar Chatterjee. Rishi Sunak promised to make AI safe. Big Tech’s not playing ball. <https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/>, 2024. [Accessed 12-09-2024].
- [19] Gabriel Mukobi. Reasons to doubt the impact of AI risk evaluations. *arXiv preprint arXiv:2408.02565*, 2024.
- [20] John Burden. Evaluating AI Evaluation: Perils and Prospects, July 2024. arXiv:2407.09221 [cs].
- [21] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, October 2017.
- [22] Marius Hobbhahn. We need a Science of Evals. <https://www.apolloresearch.ai/blog/we-need-a-science-of-evals>. [Accessed 12-09-2024].
- [23] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static AI evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, 2024.
- [24] UK Department for Science, Innovation and Technology. AI Safety Institute approach to evaluations, 2024.
- [25] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, and Henry Alexander Bradley. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. Number 1. Rand Corporation, 2024.

- [26] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.
- [27] OpenAI. OpenAI Red Teaming Network, 2023.
- [28] METR. Guidelines for capability elicitation, 2024.
- [29] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. AI capabilities can be significantly improved without expensive retraining. *arXiv preprint arXiv:2312.07413*, 2023.
- [30] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, June 2024. arXiv:2401.14446 [cs].
- [31] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI Sandbagging: Language Models can Strategically Underperform on Evaluations, June 2024. arXiv:2406.07358 [cs].
- [32] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.
- [33] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [34] Alan Chan. Evaluating Predictions of Model Behaviour. <https://www.governance.ai/post/evaluating-predictions-of-model-behaviour>, 2024. [Accessed 12-09-2024].
- [35] Scott Wiener. Sb 1047: Safe and secure innovation for frontier artificial intelligence models act., 2024.