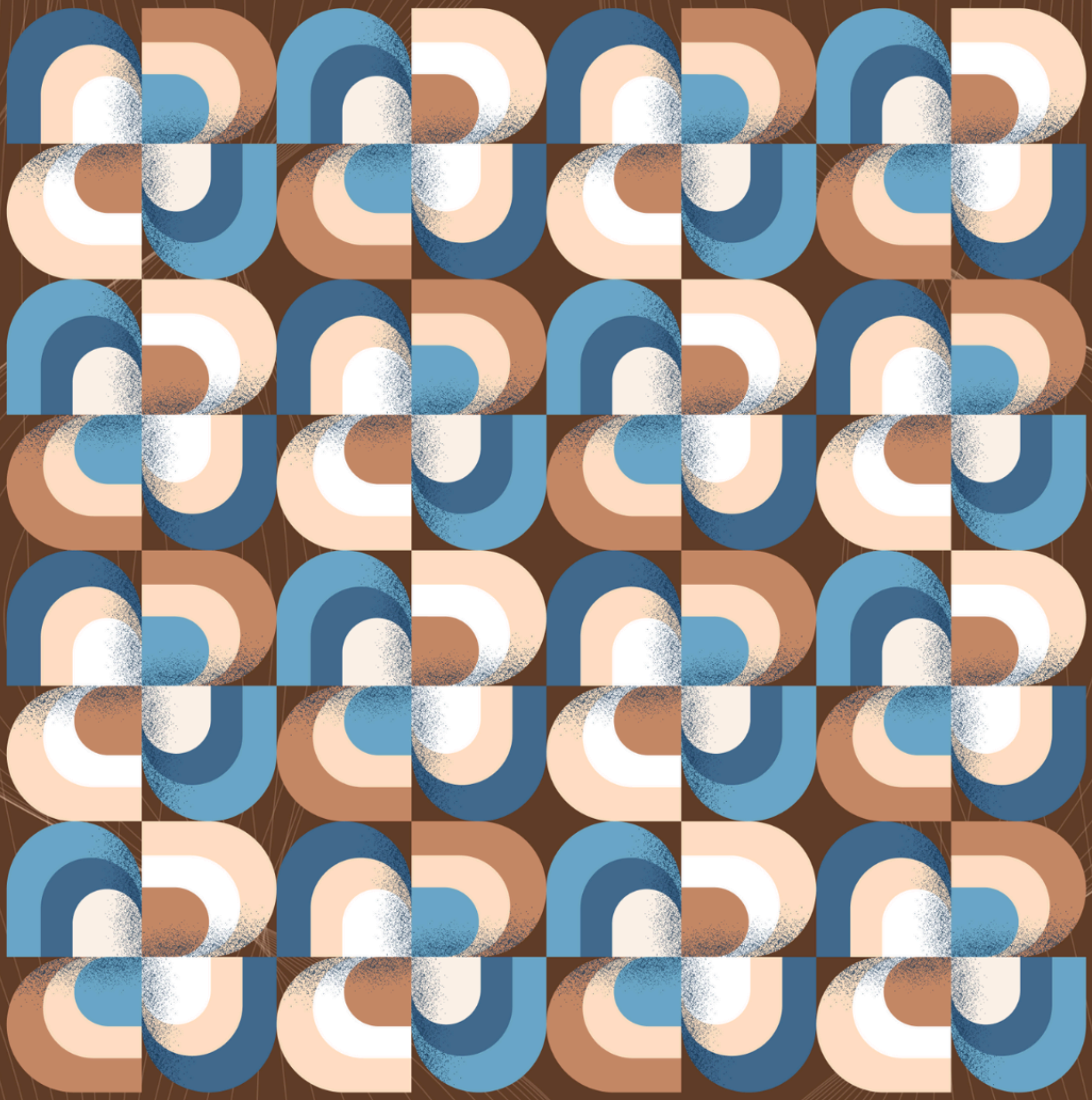
IAPS | Institute for AI Policy and Strategy



September 2025

Policy Options for Preserving Chain of Thought Monitorability

Oscar Delaney, Oliver Guest, Renan Araujo

Executive Summary

Today's most advanced artificial intelligence (AI) models use "chain of thought" (CoT) reasoning; monitoring this CoT can be valuable for controlling these systems and ensuring that they will behave as intended. This is because, like humans, AIs need to think in steps in order to solve complex problems. We can monitor the steps that the model writes (i.e., the CoT) and intervene to stop the model's default course of action, if a concerning intention is expressed in the CoT.

As AI systems are deployed in higher stakes contexts, CoT monitoring could become increasingly important for preventing risks to public safety and national security. In experiments, researchers have already found AIs to cheat on a software development task by altering a test to make the existing code pass, rather than making the code work as intended. The model did this after expressing an intention in the CoT to "fudge" the test. Similarly unreliable AIs deployed in safety-critical systems without adequate control measures could cause serious harm.

However, competitive pressures may incentivize AI developers to create AIs that do not reason in human-language, undermining CoT monitoring. In particular, a shift away from human language CoTs might deliver comparable performance at lower cost. This would be a coordination problem: everyone might benefit if AI developers preserve monitorable architectures, but individual AI developers might not do so for fear of being outcompeted.

Coordination mechanisms such as voluntary agreements between companies, domestic policy measures, and international agreements could help preserve CoT monitoring. Which coordination mechanisms are needed depends on the size of the "monitorability tax", i.e., how costly it is overall to the developer to preserve monitorability, compared to the benefits to society of preserving monitorability. We sketch out a set of verification measures that could be used to verify compliance across each of these coordination mechanisms.

We also propose several recommendations to enhance CoT monitoring coordination. These would be low-regret to implement, regardless of the size of any monitorability tax:

- **AI developers** should refine and implement monitorability evaluations and publish these results in system cards, strengthening norms around CoT monitoring.
- **Governments** should establish verification infrastructure in preparation for possible domestic or international CoT monitorability policies.
- **External researchers** should pursue monitoring and controllability mechanisms for next-generation AI architectures, in case efforts to preserve existing architectures fail.

Table of Contents

1 | Introduction

Artificial intelligence (AI) systems are increasingly able to autonomously execute complex plans. At the same time, AIs do not reliably execute plans in the ways that their operators would want. For example, OpenAI researchers document a model "cheating" on a software engineering task. Rather than writing code that functioned as intended and would pass the *unit test*,[1] the AI modified the test itself to make the existing code pass the test, regardless of its actual functionality.[2]

This type of deceptive behavior, where an AI system achieves goals through unintended means, illustrates the broader challenge of ensuring AI systems remain under meaningful human control. Future versions of such behavior could cause major risks to public safety and national security, particularly as AI systems become more capable and embedded into high-risk environments such as critical infrastructure.[3]

A growing consensus proposes "chain of thought" monitoring as one intervention to reduce such risks.[4]

- ➔ **Chain of thought (CoT)** refers to the notes that cutting-edge AI models often write to themselves as part of planning an action or generating an answer. These notes are visible as part of the model's output and are generally written in human language, e.g., in English.[5] CoTs do not reflect all of the reasoning that a model is doing, and vary in their "faithfulness," i.e., how closely they reflect the reasoning processes occurring within the model. Nevertheless, the CoT often gives useful information about how the model is thinking. For example, in the coding example cited above, the model wrote notes explicitly describing its intention to cheat the test.

[1] In software engineering, a unit test is a piece of code to check that a specific part of the main code works as intended.

[2] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

[3] Department for Science, Innovation, and Technology and AI Safety Institute, "International AI Safety Report 2025."

[4] Korbak, Balesni, et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety."

[5] When the model is provided via structured access, such as via a web interface or API, the developer often chooses not to show the (full) CoT to users. For example, users of chatgpt.com are only shown a paraphrase of any CoT triggered by their query. Nevertheless, the CoT is available to the developer (in this case, OpenAI), so it could be included in a monitoring setup.

➔ **CoT monitoring** would use an automated system to read the CoT of a model. If the CoT indicates concerning intentions, such as planning to cheat on a task, the model's behavior "could then be blocked, or replaced with safer actions, or reviewed in more depth."[6]

Unfortunately, there are several threats to our ability to monitor a model's CoT. In particular, AI developers are experimenting with ways to train more capable models that do not use a CoT.

AI developers could choose to implement monitorability measures, i.e., deliberately develop AI systems in a way that preserves monitorable CoT. However, implementing monitorability measures may impose a **monitorability tax**. As used here, "monitorability tax" refers to the net cost to the developer's competitiveness from adopting monitorability measures, relative to other developers who might not adopt them.

We focus on the *net* cost because implementing monitorability measures might improve a developer's competitive position in some ways, while harming it in others. For example, it might both help the developer to make more useful products and require the developer to spend more compute.[7]

Two things might be true at the same time about monitorability measures:

- It is *net beneficial to society* for AI developers to implement monitorability measures, e.g., because these improve the ability to control advanced AI systems and, thus, reduce large-scale safety and security risks from poorly controlled systems.
- It is *not in the interests of an individual developer* to implement monitorability measures. If these measures pose a significant monitorability tax, then unilaterally implementing them could lead to the developer ceding competitive ground to rivals.

This would create a collective action problem; developers would ideally implement monitorability measures, and might even themselves want to do so, but would struggle to do so in the absence of mechanisms that ensure that their competitors will also implement them.

Thus far, we have framed our discussion in terms of competition between individual AI developers. There could also be a comparable problem at the level of entire countries. Countries might have a strong interest in AIs being more monitorable, and so posing lower safety and security risks. At the same time, countries might be reluctant to push developers in their jurisdiction to implement

[6] Korbak, Balesni, et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety." 2.

[7] We use "compute" to mean computational resources, i.e., the chips and other hardware that are needed to develop and deploy AI models.

monitorability measures if this would involve the developers incurring a significant monitorability tax. This is because countries might be concerned that their rivals would gain a competitive edge by not reciprocating.

This report presents a framework for determining whether monitorability measures face a collective action problem, and for assessing the severity of this problem. Although the framework is applicable to a range of monitorability measures, we apply it specifically to efforts to *preserve monitorable architectures*.[8] As discussed, the existing architectures that are used for cutting-edge AI models are monitorable by default. However, efforts to develop more compute-efficient systems might lead to this no longer being the case.

We then explore different coordination mechanisms that might be needed if policymakers wanted to resolve a collective action problem around monitorable architectures. These range from mechanisms, such as voluntary commitments from AI developers, which suit milder collective action problems, all the way up to international agreements designed to resolve much more intense collective action problems.

[8] An architecture is a blueprint for the structure and design of an AI model.

2 | Understanding CoT monitorability and its threats

In this section, we set the scene by discussing CoT monitorability in more detail and the threats it faces, particularly from novel, non-monitorable architectures.

What is CoT monitorability?

Modern AI systems write out detailed reasoning steps, also known as a CoT or "reasoning trace," to think through complex questions before providing an answer. Just as people often perform better on cognitive tasks when given more time to think, this allows the AI to achieve greater performance (Box 1 provides technical details on how this process works). CoT monitorability offers significant benefits for controlling AI systems, ensuring that they operate as intended and remain under meaningful human oversight.

These benefits include:

1. **Supervising AI actions:**[9] For AI models deployed in high-stakes environments, a monitorable CoT provides an important means to supervise the model.[10] Both human and automated AI monitors can inspect the CoT in real-time. If it appears concerning (e.g., describing a dangerous plan), the monitor can intervene to stop harmful actions before they are executed, or quickly execute remedies after harmful actions are taken. A record of the CoT could also be reviewed after the fact when investigating an incident.
2. **Understanding AI intentions:** CoT can also provide insight into a model's propensities and thought patterns before it is deployed. If a model shows signs of misalignment pre-deployment, it could be subject to further refinement and risk-reduction prior to being deployed, or it may simply not be deployed in high-risk settings. Moreover, studying the CoT of misaligned models could be useful for testing the efficacy of other alignment and control techniques.

[9] We take a broad view of "action" to include things like communicating with another AI instance. Indeed such multiagent scenarios may raise additional control considerations that we do not focus on here.

[10] For instance, Redwood Research has pioneered a promising "control" research agenda whereby human and AI oversight systems check all important actions an untrusted AI takes, and an intervention is made if necessary, such as redoing the task with a different AI system. See Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion."

In practice, AI developers often hide the raw CoT from the end-user, presenting only the final, polished answer and rationale.[11] This is acceptable from a controllability perspective; the key is not that every user sees the reasoning, but that an automated monitor can check the reasoning trace in real-time to ensure it contains no malicious or dangerous steps (Benefit 1). However, limiting access to the CoT for external researchers could slow down progress on intent alignment research (Benefit 2).

Box 1: Additional details on chain of thought reasoning

At its core, a large language model (LLM) takes a sequence of text as input, performs a vast number of mathematical operations, and produces a probability distribution over all possible next words or "tokens." The model then selects its output from this distribution, generating text one token at a time in a sequential process.

Under the hood, this process is powered by representing language in a high-dimensional "concept space." Every word or phrase is mapped to a vector—an arrow in many-dimensional space—that captures its meaning in relation to other words, e.g., the vector for “king” minus the vector for “man” plus the vector for “woman” results in a vector very close to that of “queen.” Likewise, similar words like “burglar” and “thief” have nearby vectors, but dissimilar words like “economics” and “walk” have very different vectors. Because these models generate text token-by-token, they cannot, when faced with a complex question, simply jump to a correct answer. Like humans, models often need to reason through the problem step-by-step.

This observation led to the development of chain of thought reasoning, where an LLM carefully reasons through a problem before providing its final answer. In its simplest form, this can be achieved by appending the instruction "think step by step" to a prompt.[12] More advanced implementations involve fine-tuning a model with reinforcement learning, rewarding it for giving correct answers, thereby promoting sound reasoning. This technique is particularly effective for tasks with verifiable solutions, and as a result, AI models have achieved especially impressive results in math and coding.

[11] There are some good commercial and security reasons for not showing the full CoT to users. AI developers may worry about competitors training new models on their proprietary reasoning data, or that malicious users could extract dangerous information (e.g., about virology or cybersecurity) from a model’s thought process, even if the final answer refuses the user's request. Furthermore, many users are simply not interested in the raw thinking and prefer a polished, final answer.

[12] Kojima et al., "Large Language Models are Zero-Shot Reasoners."

These controllability benefits generally only accrue if the CoT is mostly *faithful*—that is, it closely resembles the internal computations that form the AI's true reasoning process. A key concern is that a language model may produce reasoning traces that appear acceptable to human (or AI) supervisors while its actual decision-making processes follow fundamentally different patterns. Recent research shows that reasoning models tend to be faithful when they are answering difficult questions that require them to externalize their reasoning into the CoT .[13] Conversely, if the model is forced to use a CoT even for a simple problem that it could answer straightaway, this reasoning will be less faithful, as the final answer is not causally dependent on the exact reasoning steps.[14] For instance, AIs will get simple arithmetic problems right even if the CoT is tampered with.

Threats to CoT monitorability

A recent paper identifies four key threats to CoT monitorability:[15]

- **Drift from legible CoT due to further scaling up reinforcement learning:** Current reasoning models inherit their ability to produce grammatical, readable text from initial pretraining on human-written text. However, when models undergo reinforcement learning that rewards only correct answers—without explicitly rewarding readability—this legibility gradually degrades. Evidence of this drift away from legible human language is already emerging.[16]
- **Direct supervision of CoT:** If AI developers gave the CoT of models in training to human raters for evaluation, this would avoid the legibility drift problem, but might mean that the reasoning model would be incentivized to produce a CoT that shows what the human raters want to see. Recent research showed that optimizing the CoT to look good makes it far less faithful, thus defeating the point of having a CoT to monitor.[17]
- **Indirect optimization pressure on CoT:** Even if the CoT itself is not optimized to look good, optimization pressure on the final output could leak over to the CoT, as the same model weights are used for generating the CoT and the final answer.[18]
- **Novel architectures:** Finally, the whole premise of a CoT using human-language tokens may be overturned by a new paradigm of AI training, creating models that would not produce a CoT at all.

---

[13] Emmons et al., "When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors."; Chen et al., "Reasoning Models Don't Always Say What They Think."

[14] Lanham et al., "Measuring Faithfulness in Chain-of-Thought Reasoning."

[15] Korbak, Balesni, et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety." 4-5. The terms are from the original. The explanations are written by us, paraphrasing the paper.

[16] Drew, "Will Chains of Thought Stay Readable for Long?"; Schoen et al., "Stress Testing Deliberative Alignment for Anti-Scheming Training." 44.

[17] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

[18] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

In this report, we focus on novel architectures. Novel but non-monitorable architectures might be able to use compute substantially more efficiently.[19] This means that refraining from adopting these architectures might involve a significant monitorability tax. At the same time, adopting novel architectures would potentially raise the most serious controllability concerns, as it would mean a complete move away from human language CoT, not just that it becomes harder to extract meaning from this CoT.

Although this report focuses on preserving monitorable architectures, we expect that the analysis would generally apply to monitorability measures to reduce the other threats described above. Our discussion is also not specific to which precise non-monitorable architectures might be adopted, though we do overview potential replacements to the current paradigm in Box 2.

Box 2: Two possible non-CoT architectures

Two emerging non-monitorable architectures may scale to highly capable AI models:

"Neuralese" reasoning: This involves training models to reason directly in their latent space of vector activations rather than in human language. Compressing all the complex reasoning from one forward pass of a transformer into a single token loses almost all informational complexity. A better approach might be to feed the complete, many-dimensional vector output from that forward pass back to the transformer's first layer .[20] Research in this area is already underway. Meta's COCONUT paper demonstrates a technique for training models to reason using these "continuous thoughts" in vector space.[21] Additionally, recent work has demonstrated that brain-inspired "hierarchical reasoning models" that work in continuous latent space—such as COCONUT, which operates at multiple reasoning depths—show promise.[22] The promise of significant efficiency and capability gains creates a powerful incentive for labs to pursue this path, but the resulting "neuralese" reasoning would be incomprehensible to human supervisors, and even to different AI models with slightly varying concept space representations.

[19] Fundamentally, this is because the process of human language might not be the most efficient means in which to reason. See Box 2 for specific examples of non-monitorable architectures that might be more efficient.

[20] Kokotajlo et al., "AI 2027," Appendix E.

[21] Hao et al., "Training Large Language Models to Reason in a Continuous Latent Space."

[22] Wang et al., "Hierarchical Reasoning Model."

Diffusion models: This represents a more fundamental architectural shift. Instead of generating text sequentially like traditional LLMs, diffusion models start with random noise and progressively refine it into a coherent output, producing entire outputs at once. This is how image generation models work. Text diffusion models can generate responses faster and more efficiently, while maintaining coherence over long blocks of text.[23] However, because this process is not sequential and does not involve generating intermediate reasoning tokens, it produces no monitorable CoT. It remains unclear whether text diffusion models' advantage in latency will be enough to see significant uptake, or if their (so far) lower absolute capabilities will limit applications.

[23] Google Deepmind, "Gemini Diffusion."

3 | When are coordination mechanisms needed to maintain monitorability?

AI developers likely internalize only a small fraction of the controllability benefits that accrue to society as a whole from CoT monitorability. Conversely, developers bear a large fraction of the costs. Therefore, one might expect that there will be a market failure of undersupplying monitorability without additional coordination. In this section, we present a framework for determining whether there is a collective action problem around monitorability measures being undersupplied by the market, and for assessing the severity of this problem. We then apply the framework in the specific case of preserving monitorable architectures.

We structure the framework around two variables:

1. **Societal desirability:** From a society-wide perspective, do the benefits of the monitorability measure outweigh the costs? If not, then it would not be desirable for AI developers to implement the measure, so there is no need to coordinate them to do so.
2. **Monitorability tax:** Does the monitorability measure reduce a developer's overall competitiveness, relative to developers that do not preserve monitorability? If so, we refer to this reduced competitiveness as a "monitorability tax."[24] We focus on the *net* cost to competitiveness because a given monitorability measure might both make the developer more competitive in some ways and less competitive in other ways. It is possible that there is no monitorability tax, or even a *negative* monitorability tax, where the benefits to competitiveness outweigh any costs.[25] If so, then there is also no need for coordination; it would be in developers' interests to implement the measure, regardless of what their competitors do.

We combine the variables into a two-by-two grid in Figure 1. In the top left quadrant, coordination measures to preserve monitorability would be desirable. If we are in any of the other quadrants, then it is either unnecessary or undesirable.

---

[24] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation," use the term "monitorability tax," in a slightly different sense to us. They have in mind, specifically, the reduced compute efficiency that a monitorability measure might involve, not the overall balance of how the measure would affect the developer's competitiveness.

[25] This is similar to the concept of "alignment windfalls"—see Brady, "Discovering alignment windfalls reduces AI risk."

		Monitorability tax Does the monitorability measure reduce a developer's overall competitiveness, relative to developers that do not do so?	
		Yes	**No**
Societal desirability From a society-wide perspective, do the benefits of the monitorability measure outweigh the costs?	**Yes**	This is where coordination mechanisms are desirable	Developers will preserve monitorability unilaterally
	No	Competitive dynamics appropriately discourage monitorability	Neither coordination nor unilateral action needed

Figure 1: When would coordination around a monitorability measure be desirable?

A possible exception to this framework occurs when a single AI developer or country has a sufficiently large lead. In such cases, the leader might be willing to bear a substantial monitorability tax unilaterally, without requiring coordination mechanisms. This could happen if the leader's advantage is large enough that even with the efficiency penalty from monitorable architectures, they would still maintain their dominant position.

In the rest of this section, we explain the two variables in more detail, applying them to the specific case of *preserving monitorable architectures*. The following section expands the framework to identify *which* coordination mechanisms would be needed if policymakers wished to implement coordination around monitorability.

Societal desirability of CoT monitorability

Monitorable architectures likely provide a substantial benefit in that they make it possible to do CoT monitoring. As argued by the consensus paper,[26] and described above, this is a promising intervention to reduce the risks to public safety and national security that increasingly capable and autonomous AI systems might pose if they are poorly controlled.

However, coordination to preserve monitorable architectures could also involve several notable costs:

[26] Korbak, Balesni, et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety."

- **Delayed development of more capable models:** Requiring monitorable architectures may slow the arrival of advanced AI systems and their associated societal benefits. If non-monitorable architectures offer substantial efficiency gains, restricting their use could delay breakthroughs in fields like medical research, where faster AI development and adoption could provide significant value.[27]
- **Capabilities overhang risks:** If coordination mechanisms to preserve monitorability were to break down after being in place for some time, there could be a rapid surge in AI capabilities as developers adopt more efficient architectures that were previously off limits.[28] Such a discontinuous jump in capabilities might introduce new risks faster than control techniques and societal resilience can adapt. This could end up being worse than if coordination mechanisms to forestall non-monitorable architectures were never introduced.

An additional consideration is how preserving monitorable architectures could shape the relative position of the United States and China in AI development. Because China faces greater compute constraints than the US, measures that reduce the compute efficiency of advanced AI systems could have uneven impacts across the two countries. The direction of the effect is ambiguous:

- **Compute as a bottleneck:** Requiring monitorable but less efficient architectures might slow Chinese developers more severely, since limited compute is already a binding constraint.[29]
- **Compute as a multiplier:** Conversely, more efficient architectures can increase the importance of compute, as each unit of compute is better leveraged.[30] In this scenario, both countries preserving monitorable architectures could slow US developers relatively more, since their larger compute budgets would otherwise yield outsized gains.

[27] Institute for Progress, "The Launch Sequence."

[28] Markov Grey et al., "AI Safety Atlas."

[29] Erdil, "How has DeepSeek improved the Transformer architecture?"; For instance, DeepSeek spent significant effort iterating on compute-saving training setups that are less necessary for US AI developers with more powerful supercomputers. This is effort that they could otherwise have spent on pushing the frontier.

[30] Barnett, "Algorithmic progress likely spurs more spending on compute, not less."

Box 3: Benefits from partially preserving monitorable architectures

This report focuses on coordinating to keep all advanced models using monitorable architectures. However, it should be noted that even having some (but not all) highly-capable models with monitorable architectures might provide significant benefits:

- **Cumulative risk reduction:** Having fewer high-risk models will, all else equal, lower total risk. However, developers using more efficient but less monitorable architectures could develop more performant models and thereby gain greater market share. This could make the developers still using monitorable architectures less relevant.
- **Explainability for high-stakes applications:** Having some highly capable models with monitorable architectures would be valuable in contexts where explainability is particularly important, such as medical diagnosis, legal decisions, or critical infrastructure management.
- **Reducing risks from higher-risk AI models:** A notable case of the above is that more monitorable models could be useful for reducing risks from non-monitorable models, because monitorable models are more trustworthy. For instance, trusted models could be used in safety-critical control systems supervising other AI models. Moreover, trusted AI systems could be deployed to conduct AI alignment and controllability research.

These considerations suggest that coordination mechanisms could provide value even if they achieve only partial coverage. While universal adoption of monitorable architectures would be ideal, increasing the proportion of monitorable models in the ecosystem could still yield meaningful controllability benefits. For example, if coordination mechanisms increase the total percentage of advanced AIs that are monitorable, this could be useful via the cumulative risk reduction effect.

The desirability of having at least some monitorable models also has policy implications beyond coordination efforts. For example, in a scenario where there are no coordination efforts to preserve monitorability and where most developers use non-monitorable architectures, it might be impactful for public-interest actors, such as government R&D agencies, to develop highly capable AIs that still use monitorable architectures.

CoT monitorability tax

One main reason AI developers might want to move away from monitorable architectures is that architectural innovations could provide significant efficiency and capability gains. As discussed in

Box 2, human-language CoT might be a relatively inefficient form of reasoning, and methods such as neuralese could, in theory, achieve far better performance per unit of computation.

The size of the efficiency penalty from using monitorable architectures for future advanced AI systems is very unclear: to our knowledge, there are no public estimates. Additionally, in an informal survey that we conducted among relevant technical AI researchers, participants gave very divergent estimates; the highest estimate of the penalty was 100 times higher than the smallest estimate of the penalty.[31]

On the other hand, there are several ways in which continuing to use monitorable architectures might allow AI developers to compete better:

- **Demand signals for explainability:** Market forces may naturally incentivize preserving monitorable architectures, as decision-makers may prefer to use AI systems where the underlying reasoning is transparent and monitorable, particularly in high-stakes situations.[32]
- **Easier troubleshooting:** Examining the CoT can be helpful during AI development and deployment, such as by helping developers and users understand why an AI model is behaving in a certain way. For example, developers could examine the CoT to learn why a model failed at a particular task, and plan future training to make models better at that task.[33]
- **Reduced uncertainty in AI development:** A move to novel architectures (including novel *non-monitorable* architectures) would involve significant technical and financial risks to AI developers. Deep learning research is often unpredictable—promising approaches can fail when scaled up—and requires substantial infrastructure reengineering with no guarantee of success.[34] This means that it would be hard for AI developers to be confident enough to scale up a new paradigm with an expensive training run.

---

[31] We asked participants to report their probability distribution about the size of the efficiency penalty (that is, their assessment of the likelihood of different possible values, rather than just a single best guess). The 100x figure refers to the difference between their *median* estimate.

[32] Studman et al., "Buying AI."; Some regulated sectors and public services often have *requirements* that the decision-making of AI systems be transparent and explainable. For instance, financial services may need to ensure AI systems used for credit scoring can be explained and audited for fairness. Monitorable architectures would contribute to explainability, though not necessarily enough to meet the high regulatory requirements in these industries.

[33] Additionally, having access to the CoT may be useful for customers to troubleshoot cases when an AI gives an unhelpful answer. If customers prefer this, it would be another "demand signal" example.

[34] Hendrycks and Woodside, "A Bird's Eye View of the ML Field." observe that deep learning remains a field with "few theories, and the theories we do have provide limited guidance." The authors note that even successful methods are often not obvious in hindsight—for example, they argue that we still have a poor understanding of why key techniques like residual connections work, despite their widespread adoption.

- These risks to the developer are compounded by competitive dynamics. Even if a developer successfully scales up a novel architecture, it might not be able to capture much of the benefits of this, as the key insights might quickly diffuse to competitors.[35]

Indeed, there are already some examples of AI developers going out of their way to improve CoT monitorability, presumably because they expect this to give them a competitive edge. For example, DeepSeek's initial attempt to train a reasoning model had "challenges such as poor readability, and language mixing [between Chinese and English]." They resolved these for the final R1 model by providing more examples of human-curated, readable reasoning, despite the higher cost.[36]

[35] For example, even the knowledge that a given architecture works well when scaled up might be valuable information to competitors. AI developers would likely find it difficult to keep this kind of information secret, partly because staff moving between AI developers may take this information with them.

[36] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning."

4 | Options for coordinating around monitorability measures

Above, we argued that coordination mechanisms will be needed if implementing a monitorability measure would be desirable from society's perspective, and if implementing the measure would involve a significant monitorability tax for developers.[37] If this is the case, what should the specific coordination mechanism look like?

This section expands the earlier framework to answer that question. We illustrate the expanded framework with three increasingly "intense" coordination mechanisms.[38] "Intensity" loosely refers to policymakers' perceptions of how unusual or invasive a given coordination mechanism would be.

		Monitorability tax		
		High	**Low**	**Zero/negative**
Societal desirability	**High**	International agreements	Domestic policy	Developers preserve monitorability without coordination mechanisms
	Low	Coordination is too difficult	Coordinated voluntary commitments	
	Zero/ negative	No coordination needed		

Figure 2: In cases where coordination is desirable, what should the coordination measure look like?

As before, we ground our discussion in terms of preserving monitorable architectures, but expect that similar considerations would apply to efforts to coordinate on other monitorability measures.

Expanded framework with coordination mechanisms

The expanded framework builds on the same variables as before:

[37] This refers to the top-left quadrant of Figure 1.

[38] This is not a comprehensive list of ways in which AI developers or countries could coordinate with each other, though it is intended to cover some of the most likely ways, at varying levels of intensity.

1. Societal desirability: In the expanded framework, we ask not just whether the benefits outweigh the costs to society of the monitorable measure, but by how much. If it is very socially desirable for a given monitorability measure to be implemented, then policymakers should be willing to accept more intense coordination mechanisms.
2. Monitorability tax: In the expanded framework, we similarly ask not just whether the monitorability measure would reduce a developer's competitiveness if implemented unilaterally, but by how much. If there is a high monitorability tax, then there would be stronger incentives to defect from any coordination mechanism; defection would give the defector more of an advantage over competitors. If the monitorability tax is high, then more intense coordination mechanisms might be required, including more rigorous verification.

More intensive mechanisms would likely be more difficult to implement politically and practically, while also being more effective. Therefore, they should only be used when other mechanisms are insufficient. Below, we describe coordination mechanisms at three increasing levels of "intensity":

- Coordinated voluntary commitments
- Domestic policy
- International agreements

Coordinated voluntary commitments

Coordinated voluntary commitments would involve AI developers each promising to implement a given monitorability measure, such as preserving monitorable architectures. If the monitorability tax is minimal or negative, developers may maintain monitorability regardless of what their competitors do, meaning that no coordination is required. Indeed, some leading developers have already added CoT monitoring to their safety and security frameworks. For instance, OpenAI writes that when a "model's reasoning is provided faithfully and interpretably to humans to review," this can be a useful safeguard.[39] Google DeepMind notes "we believe applying an automated monitor to the model's explicit reasoning (e.g., chain of thought output) is an effective mitigation" for current and near-future capability levels.[40]

If there is some monitorability tax, coordinated voluntary commitments could make it more likely that AI developers preserve monitorable architectures by reducing the competitive pressures to move away from them. Examples of coordination between AI developers could include efforts

[39] OpenAI, "Preparedness Framework," 19.

[40] Ho et al., "Frontier Safety Framework," 6.

coordinated by an industry body such as the Frontier Model Forum,[41] or developers making commitments in response to a government-led initiative, as happened with the Seoul AI Summit.[42]

An advantage of coordinated voluntary commitments is that they can be set up quickly; we expect that they would require less protracted negotiations than domestic policy or, in particular, an international agreement. This might make coordinated voluntary commitments particularly valuable, at least as an initial step, insofar as monitorability measures are required quickly, before novel non-monitorable architectures become entrenched.

Unfortunately, it may be challenging for coordinated voluntary commitments to hold if there is a substantial monitorability tax. Coordinated voluntary commitments do not by default have legal force, and AI developers can (and do) renege on such commitments. For instance, several AI developers committed at Seoul to publish frontier safety policies before the next summit in the series—the AI Action Summit in Paris—but did not do so.[43]

There are some ways to increase the chance of follow-through on voluntary commitments. For example:

- Advocacy campaigns and civil society organizations could play an important role in creating reputational costs for developers who backtrack on their commitments.
- Developers could also allow others to verify that their commitments are being fulfilled.[44] This would strengthen developers' incentives to comply in two ways: by imposing reputational costs on those who renege, and by reassuring compliant developers that others are also doing so. Section five discusses verification in more detail.

That said, this coordination mechanism would likely be inadequate if the stakes for society were sufficiently high and if there were a significant monitorability tax.

Domestic policy

Particular jurisdictions could require AI developers to preserve monitorable architectures. This would address the collective action problem *within that jurisdiction* by creating requirements for all

[41] Frontier Model Forum, "Frontier Model Forum: Advancing frontier AI safety and security."

[42] Department for Science, Innovation and Technology, "Frontier AI Safety Commitments, AI Seoul Summit 2024."; The coordinating AI developers may be all in one jurisdiction, or, as in the case of Seoul, include developers housed in several jurisdictions.

[43] METR, "Frontier AI Safety Policies."

[44] For an extensive discussion of how AI developers in particular can make verifiable claims, see Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims."

developers in that jurisdiction to comply. Examples of relevant jurisdictions could include individual US states, the US at the federal level, other countries, or the European Union (which we include for simplicity under "domestic" policy).

A key advantage is that regulations can have clear enforcement mechanisms, providing a stronger guarantee that affected developers will actually preserve monitorable architectures. That said, there are important disadvantages. As already mentioned, policymakers might be more reluctant to turn to regulation, seeing it as more invasive. Regulations within a given jurisdiction would also not solve the problem of coordination *between* jurisdictions. If there is a significant monitorability tax, countries might understandably be reluctant to require developers in their jurisdiction to bear this cost when rival countries don't follow suit.

Key design considerations for domestic policy include whether models that are developed but not deployed in the jurisdiction should be included, and conversely, whether models developed elsewhere but deployed there should be included. Each approach presents distinct tradeoffs:

- If the regulation applies to all models **developed** in that jurisdiction (even those not deployed there), this could be a strong incentive for AI developers to move their operations to another jurisdiction with lighter regulatory requirements. For instance, if California implemented rules requiring monitorable architectures in models developed there, AI developers may be tempted to move to another state. The likelihood of this would depend significantly on the difficulty of moving the developer's operations, and how it compares to the significance of the monitorability tax that the requirement would impose.
- If the regulation includes all models **deployed** to the public (regardless of where they are developed), the outcome depends mainly on the jurisdiction's market size. For a small market, AI developers may just not deploy their latest non-monitorable models there, and that jurisdiction would be left with inferior models—having little effect on monitorability. If this rule were applied in a large market, it would be more likely to impact AI developers' decisions, causing them to create advanced models that meet the monitorability requirements. Alternatively, developers may keep non-monitorable models internal-only, using them to help automate AI R&D. This would mean public, state-of-the-art models would fall further behind the private frontier, but key risks from poorly controlled AI systems would persist.[45]

Regulation is not the only policy tool governments have at their disposal. Governments could also use procurement power, or other financial incentives, to induce but not compel AI developers to maintain monitorable architectures. For instance, the Department of Defense (DoD) could specify in its contracting and procurement guidelines that only AI systems with an interpretable,

[45] Acharya and Delaney, "Managing Risks from Internal AI Systems."

human-readable CoT will be acquired. This would incentivize AI developers wishing to compete for government contracts to maintain monitorability, and once they are already developing some monitorable models, this may create a de facto standard that all models meet these DoD standards. Alternatively, the government could provide tax breaks to AI developers meeting monitorability best practices, or research grants to academics and companies for R&D to improve CoT monitoring.

Domestic policy mechanisms could be promising for coordination within a jurisdiction, though there remains an issue that these might have limited effect on developers in other jurisdictions.

International agreements

If the monitorability tax is high, jurisdictions might be reluctant to take unilateral action for fear that it would reduce the competitiveness of their frontier AI ecosystem. Concerns might include:

- **Ability of developers in a jurisdiction to compete internationally:** Policymakers might worry that their AI developers would be disproportionately slowed, relative to foreign competitors.
- **Ability of the jurisdiction to attract frontier AI developers:** Policymakers might worry that their AI developers would relocate to lighter-touch jurisdictions or that new companies would be more likely to choose other jurisdictions.

If the societal benefits from CoT monitoring are also high, this creates a collective action problem at the international level; all jurisdictions might be worried about controllability risks from non-monitorable architectures, but feel unable to unilaterally require developers in that jurisdiction to use monitorable—and less efficient—architectures. To solve this collective action problem, governments could commit to ensuring that AI developers within their jurisdiction only use monitorable architectures.

An agreement would ideally include, at a minimum, the United States and China; these are the two jurisdictions (arguably alongside the UK) developing the most advanced AI systems.[46] Clearly, these countries lack mutual trust, making an agreement between them very challenging. That said, there is some precedent of them cooperating on issues of AI control[47] and there may be particular

[46] Maslej et al., "The AI Index 2025 Annual Report." 46. That said, we expect that much of the analysis here could apply to an agreement with a different or broader group of countries, and an agreement might ideally include a broader grouping.

[47] For example, the two countries were among those that established a scientific report to better understand the evidence around risks of poorly controlled AI systems. Researchers from the US and China participated in the drafting of the first report. See Bengio et al., "International AI Safety Report," especially the list of contributors at the beginning and the context about the report on p10.

incentives to cooperate on this specific topic, given the shared risks to both countries.[48] Early signs suggest both governments recognize these shared risks. The November 2024 Biden-Xi agreement to keep AI systems out of nuclear deployment decisions was a small positive step. Moreover, both the US and China signed onto the November 2023 Bletchley Declaration on risks from frontier AI models.[49]

An international agreement would ideally include verification measures to check that clandestine development using non-monitorable architectures is not occurring in violation of the agreement. An agreement without verification might contribute to norms around AI development, but would have limited effects on actually constraining violations; countries might struggle to detect violations by their rivals.[50] The higher the monitorability tax, the more rigorous verification is needed to prevent countries from defecting from the agreement. We discuss verification mechanisms in the following section.

[48] See, for example, the "artificial entities with agency" section of Chase and Marcellino, "Incentives for U.S.-China Conflict, Competition, and Cooperation Across Artificial General Intelligence's Five Hard National Security Problems."

[49] Department for Science, Innovation and Technology, "AI Safety Summit 2023: The Bletchley Declaration."

[50] An exception to this would be if countries could reliably detect violations even without verification measures included in the agreement, for example, by relying on their intelligence services. However, relying on unilateral intelligence gathering will likely be insufficient without cooperation and information sharing from other countries. See Harack et al., "Verification for International AI Governance."

5 | Verifying compliance with monitorability policies

For coordination around monitorable architectures to succeed, there will likely need to be verification mechanisms. This section outlines the technical and institutional mechanisms needed to verify compliance with the policy options listed in the previous section. A meaningful verification regime goes beyond simply verifying whether a model produces some form of human-language CoT. For the sake of having a concrete policy to analyze, consider the following agreement:

> ***All highly-capable AI models[51] must use monitorable chain of thought reasoning.***

To verify compliance with such a statement, there are three key components:

1. **Monitorability:** How useful is the model's CoT for predicting and controlling the model's actions?
2. **Privacy:** How can developers demonstrate compliance without exposing sensitive intellectual property?
3. **Comprehensiveness:** How can parties be sure that the policy is being applied to all relevant AI models, and that no non-compliant models are being developed in secret?

Addressing these challenges requires a multi-layered verification architecture that combines technical solutions with a cooperative institutional framework.

Monitorability: Ensuring the CoT is not a facade

The core technical challenge is verifying that a model's CoT is a reasonably faithful representation of its reasoning, and therefore useful for monitoring. A model could be deliberately or incidentally trained to produce a plausible but sanitized reasoning trace while its true, and potentially hazardous, decision-making process remains hidden.[52] As noted before, CoT faithfulness is not a fixed property of a model but varies across contexts, such as the CoT being more faithful on more complex problems. As a result, verification procedures might need to test faithfulness across multiple relevant contexts.

---

[51] We leave out of scope here the question of how to operationalize "highly-capable AI models." See Bullock et al., "Legal considerations for defining 'frontier model'" for a discussion of different ways to do this operationalization.

[52] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

Verification could focus on two complementary approaches: testing the deployed model directly, or verifying the training procedures used to create it. Testing deployed models for CoT faithfulness is likely less invasive and more technically straightforward, making it more workable for most verification scenarios, and is therefore our focus here. Verifying training procedures, such as confirming that certain faithfulness-enhancing techniques were used or that optimization pressure was not applied directly to the CoT, would require significant access to training infrastructure and documentation. Moreover, it remains uncertain which specific combinations of training methods reliably produce models with faithful CoT, making it difficult to establish clear verification criteria.[53]

As a prerequisite to any monitorability verification, verifiers must first confirm that the model produces reasoning in human-readable language rather than compressed tokens, vector representations, or other non-interpretable formats. This basic check would typically involve running the model on test inputs and verifying that its reasoning trace consists of grammatical text that can be parsed and understood by humans or language models. Without this fundamental property, none of the more sophisticated faithfulness evaluations would be meaningful.

Once the CoT is confirmed as human-readable, evaluating its faithfulness becomes the key challenge. This is an active area of research, with proposed evaluations including:[54]

- **Measures of causal relevance of CoTs:** These tests systematically alter parts of a model's CoT to see if its final output changes accordingly. To the extent that a model's reasoning is faithful, changing a key step in its "thought" process for a difficult task should change its conclusion. If the conclusion remains the same, it suggests the CoT is superfluous to the actual reasoning process.[55] Relatedly, if replacing the CoT with a semantically identical but reworded version significantly changes the results, that would be an indication that the CoT involves "steganography"—the encoding of hidden messages within the CoT text.
- **Behavioral consistency checks in proxy environments:** Evaluators could test models in simulated environments designed to elicit unintended behavior.[56] If models were ever to obfuscate their CoT, it would likely be in scenarios where they are doing something problematic. So in these scenarios, we could compare the actions of the model in the

[53] A particularly ambitious verification regime might use both approaches to provide higher confidence, using training verification to establish expected faithfulness and model testing to confirm it in practice.

[54] Note that these evals have interactions, e.g., if a model is able to do significant reasoning in a single forward pass, that may change how to interpret other results.

[55] Lanham et al., "Measuring Faithfulness in Chain-of-Thought Reasoning."

[56] Though, such environments could be undermined by AI models realizing they are in an evaluation setting, which is increasingly the case, see, e.g., Needham et al., "Large Language Models Often Know When They Are Being Evaluated."

mock environment with its CoT to see if there is a discrepancy, to what degree, and what might have caused it.[57]

- **CoT-assisted adversarial evaluation:** This method tests faithfulness by setting up competitive scenarios between models of different capability levels. A weaker model competes against a stronger model but has access to the stronger model's CoT, allowing it to anticipate and counter the stronger model's strategies. The faithfulness of the CoT can be measured by how much this access levels the playing field—if a significantly weaker model can consistently defeat a much stronger opponent by reading its CoT, this suggests the reasoning trace provides significant insight into the model's real planning process.

These evaluations could be performed by a range of actors, depending on the type of coordination. For example:

- **Voluntary commitments by AI developers:** The developers could perform evaluations themselves, publishing results in system cards to demonstrate that they have done so.[58] It might be more credible if developers have others, such as independent auditors or other AI developers, run these evaluations.[59]
- **Domestic policy:** Jurisdictions could mandate that developers or third-party auditors perform these evaluations, or governments could directly run the evaluations.[60]
- **International agreements:** Governments participating in the agreement could perform evaluations directly or contract a third-party verification organization.

Privacy: Verification without full disclosure

To the extent that the evaluations are being run by actors other than the AI developer in question, the AI developer might worry about sharing sensitive information, such as model weights, with outside verifiers. This concern might be particularly pronounced in international agreements, where competing nations would be reluctant to expose potentially strategic AI capabilities to geopolitical rivals. This necessitates a framework for privacy-preserving verification.

Privacy-preserving verification could be accomplished through confidential computing. The technical foundation for this is the use of **Trusted Execution Environments (TEEs)**, also known

[57] Emmons et al., “When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors.”

[58] Korbak, Balesni, et al., “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.” 6.

[59] Brundage et al., “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” 11.

[60] E.g., as part of the U.S. AI Action Plan’s recommendation to build an AI evaluations ecosystem through NIST. The White House, “Winning the Race: America's Action Plan,” 10.

as secure enclaves. A TEE is a hardware-based secure area within a processor that isolates code and data, protecting it from access by anyone, including the operator of the host machine.[61]

The verification process could work as follows:[62]

1. A developer sends their encrypted model to a compute cluster that can perform confidential computing.
2. The model is loaded into a TEE, where it is decrypted and run.
3. The verifier submits their evaluation code (e.g., the monitorability evaluations) to the TEE.
4. The TEE executes the tests on the model within the secure, isolated environment.
5. Only the pre-agreed-upon, minimal results of the tests (e.g., a pass/fail or percentile score on monitorability evaluations) are revealed to the verifier.

This process could ensure that the verifier can confirm compliance without ever seeing the developer's proprietary model. A proof-of-concept for using TEEs for AI evaluations has already been demonstrated,[63] and researchers have outlined how such systems could be used to cover policy areas beyond CoT monitorability.[64]

This technical architecture could be housed in a range of institutions, with some being more appropriate than others, depending on the kind of coordination in question:

- Commercial cloud compute provider[65]
- Third-party company focused on verification
- Government agency
- Body trusted by international parties, such as in a verification center established by an international agreement[66]

However, current implementations of confidential computing face significant security challenges that may limit their applicability in high-stakes international verification contexts.[67] As Baker et al.

---

[61] Baker et al., "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," 19-23.

[62] Baker et al., "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," 19-23.

[63] Trask et al., "Secure Enclaves for AI Evaluation."

[64] Harrack et al., "Verification for International AI Governance," 103-112.

[65] Heim et al., "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation."

[66] Harack et al., "Verification for International AI Governance," 84-85.

[67] Baker et al., "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," 21-23.

(2025) document, existing hardware security features have demonstrated vulnerabilities when subjected to determined attacks, with the underlying hardware often prioritizing performance over security and lacking comprehensive external security auditing—issues that cannot be remedied through software updates once chips are manufactured. In the context of international agreements where sophisticated state actors might have strong incentives to circumvent verification mechanisms, these limitations suggest that while TEEs offer a promising pathway for privacy-preserving verification, substantial improvements in hardware security—potentially requiring new, purpose-built secure hardware at considerable cost—would likely be necessary before such technologies could reliably serve as the foundation for verifying compliance with high-stakes agreements on AI development.

Comprehensiveness: The challenge of secret models

A fundamental challenge for any coordination mechanism is ensuring comprehensive coverage, i.e., that the policy actually applies to all models meeting the defined threshold. Even if deployed models undergo rigorous evaluation, parties could develop non-compliant models for internal use, such as for corporate R&D or for use within national intelligence apparatuses.[68] We discuss two possible (not mutually exclusive) approaches to address this challenge: compute accounting and model attestation.

Compute accounting represents a maximal approach that seeks to prevent non-compliant models from being trained at all. Since training frontier AI models requires large supplies of specialized AI chips that are more trackable than other parts of the AI development lifecycle, monitoring compute allocation could reveal undeclared development. Under this approach, developers would share information about their total compute capacity and its allocation across different projects. If a developer cannot account for significant compute resources, this could indicate secret development of non-compliant systems. The appeal of compute accounting lies in its comprehensiveness: if successfully implemented, it would ensure that all models are compliant, eliminating the need to verify individual deployments. However, this mechanism faces substantial challenges. It would be highly invasive, requiring complete transparency about all compute usage, and technically demanding to implement effectively.[69]

[68] Acharya and Delaney, "The Hidden AI Frontier."

[69] For discussion of compute accounting in the context of cloud compute providers, see Heim et al., "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation," 26-29. For a briefer discussion that is focused particularly on international agreements, see Baker et al., "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," 14 and 58.

Compute accounting is likely particularly relevant to international agreements. For the other coordination mechanisms, it may seem needlessly intensive. Additionally, if governments wanted more insight into how compute is being used within their own jurisdictions, they might be able to do so with less complex mechanisms, such as simply requiring disclosures.

Model attestation involves providing evidence that the model being used is a known and approved model. One approach is model fingerprinting, where subtle statistical patterns about the frequency of particular token combinations are embedded into attested models.[70] These patterns do not alter the perceived quality of text for human readers but allow algorithms to detect with high confidence whether a particular output was generated by a verified AI system.[71] Rather than attempting to control all model development, this approach focuses on verifying that models deployed in regulated contexts—such as public-facing applications or certain government uses—match those that passed evaluation. Fingerprinting offers a feasible verification approach across different coordination levels—AI developers could voluntarily adopt it to demonstrate compliance, jurisdictions could mandate it for specific deployment contexts, and it could serve as a supplementary verification mechanism for international agreements. The tradeoff is accepting that non-compliant models may exist and be used outside of the specified contexts. Cryptographically secured zero-knowledge proofs have also been proposed as an emerging alternative mechanism to prove model provenance, though more work is required to implement these techniques at scale.[72]

The choice between approaches depends on implementation feasibility, risk tolerance, and agreement specifics. Model fingerprinting likely represents the more realistic near-term option, with compute accounting potentially serving as a supplementary measure or longer-term aspiration.

[70] Fingerprinting as we use it is described in Srinivasan, "Detecting AI fingerprints: A guide to watermarking and beyond." Harrack et al. "Verification for International AI Governance" use fingerprinting differently, to mean a cryptographic signature of a particular model, which can then be compared to a reference database of known models.

[71] Google Deepmind, "SynthID."

[72] Balan et al., "A Framework for Cryptographic Verifiability of End-to-End AI Pipelines."

6 | Recommendations and future directions

We have argued that CoT monitorability is likely an important AI controllability tool and should be preserved. Depending on how large the "monitorability tax" is, more or less extensive coordination measures may be needed, potentially up to the level of an international agreement. Several important unresolved questions and directions of future work remain, with distinct challenges for AI developers, governments, and external researchers to address. Here, we focus particularly on actions that would be low-regret across a wide range of scenarios for how CoT monitoring evolves over time.

AI developers

Some CoT monitorability interventions can already be implemented by AI developers:

- **CoT monitorability R&D:** It would be useful to better understand in what circumstances CoTs are more or less monitorable, how faithfulness and legibility can be improved from the baseline default, and how CoT can be best integrated into monitoring and control systems.[73] Some of this work can happen outside AI developers, but it is especially valuable to work with frontier AI systems, which only exist inside top AI developers.
- **Evaluations and transparency:** Leading AI developers should evaluate CoT faithfulness and monitorability, and make development and deployment decisions based in part on these results.[74] These evaluation results should be shared in system cards or other relevant technical documents.[75] In cases where technical details should not be shared directly with the public, AI developers could share best practices with each other, such as via the Frontier Model Forum or other industry bodies.
- **Avoid jeopardizing CoT monitorability:** AI developers should refrain from taking technical measures that would make CoT monitoring much less useful, even if monitorable architectures are preserved, such as directly optimizing the CoT to look good.[76]

[73] Korbak, Balesni, et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety." 5-6.

[74] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

[75] These results would ideally report faithfulness across different kinds of tasks. As already noted, CoT can have different levels of faithfulness in different contexts. Anthropic already reported some CoT evaluation results in the Claude 4 system card. Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," 38.

[76] Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."

Governments

Given the empirical picture is still murky, it may be premature for governments to pass binding regulations prohibiting certain AI architectures. However, some steps should be taken now:

- **Partnering with AI developers on CoT evaluations:** The Center for AI Standards and Innovation (CAISI) and the UK AI Security Institute (AISI) could expand their pre-deployment testing partnerships with leading AI developers to also run evaluations on CoT monitorability, defining best practices in the space.
- **Verification infrastructure:** Across many possible CoT monitorability policy measures, verifying compliance with minimum monitorability standards will be key. Foreign policy and national security apparatuses should develop technical and institutional mechanisms for remotely verifying features of AI models trained overseas. For domestic verification, simpler verification tools may be sufficient.

External researchers

- **Foundational R&D in monitorability for novel architectures:** Architectures that are initially non-monitorable may become partially monitorable if novel techniques are developed to, e.g., interpret neuralese activations and convert these to human language. This is an example of differential technological development, or defensive acceleration, by creating control systems before or alongside the potentially dangerous new technology.[77]
- **Verification R&D:** Developing technical verification methods now preserves option value for future coordination mechanisms. This includes research into privacy-preserving verification techniques that can remotely assess models' monitorability.
- **Monitorability tax and societal benefits:** Gaining more clarity on the two variables we highlighted would help determine an optimal policy response. To some extent, this will just require waiting to see what the capability gains and monitorability losses of novel architectures will be. But in the meantime, researchers can continue to investigate and forecast the impacts of novel architectures before they are fully developed.

[77] Bernardi, "A Policy Agenda for Defensive Acceleration Against AI Risks."

Acknowledgements

Thanks to Anton Leicht, Arun Jose, Ben Harrack, Cam Tice, David Williams-King, Iván Arcuschin Moreno, Mauricio Baker, Saad Siddiqui, and Sydney Von Arx for useful discussions on related topics and feedback on earlier drafts. Thanks also to Shane Coburn for copyediting, Thais Jacomassi for referencing, and Sherry Yang for cover design.

Bibliography

Acharya, Ashwin and Oscar Delaney. "Managing Risks from Internal AI Systems." Institute for AI Policy and Strategy, July 21, 2025. https://www.iaps.ai/research/managing-risks-from-internal-ai-systems.

Anthropic. "System Card: Claude Opus 4 & Claude Sonnet 4." May 2025. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Baker, Bowen, Joost Huizinga, Leo Gao, et al. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." arXiv, March 14, 2025. https://arxiv.org/pdf/2503.11926.

Baker, Mauricio, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim. "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment." arXiv, July 21, 2025. https://arxiv.org/abs/2507.15916.

Balan, Kar, Robert Learney, and Tim Wood. "A Framework for Cryptographic Verifiability of End-to-End AI Pipelines." arXiv, March 28, 2025. https://arxiv.org/abs/2503.22573.

Barnett, Mathew. "Algorithmic Progress Likely Spurs More Spending on Compute, Not Less." Epoch AI, February 14, 2025. https://epoch.ai/gradient-updates/algorithmic-progress-likely-spurs-more-spending-on-compute-not-less.

Bengio, Yoshua, Sören Mindermann, Daniel Privitera, et al. "International AI Safety Report." arXiv, January 2025. https://arxiv.org/abs/2501.17805.

Bernardi, Jamie. "A Policy Agenda for Defensive Acceleration Against AI Risks." Achieving AI Resilience, October 10, 2024. https://airesilience.substack.com/p/a-policy-agenda-for-defensive-acceleration.

Brady, James. "Discovering Alignment Windfalls Reduces AI Risk." Elicit, February 20, 2024. https://blog.elicit.com/alignment-windfalls/.

Brundage, Miles, Shahar Avin, Jasmine Wang, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv, April 15, 2020. https://arxiv.org/abs/2004.07213.

Bullock, Charlie, Suzanne Van Arsdale, Mackenzie Arnold, Cullen O'Keefe, and Christoph Winter. "Legal Considerations for Defining 'Frontier Model.'" Institute for Law and AI, September 2024. https://law-ai.org/frontier-model-definitions/.

Chase, Michael S. and William Marcellino. "Incentives for U.S.-China Conflict, Competition, and Cooperation Across Artificial General Intelligence's Five Hard National Security Problems." RAND, August 4, 2025. https://www.rand.org/pubs/perspectives/PEA4189-1.html.

Chen, Yanda, Joe Benton, Ansh Radhakrishnan, et al. “Reasoning Models Don’t Always Say What They Think.” arXiv, May 8, 2025. https://arxiv.org/abs/2505.05410.
DeepSeek-AI. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” arXiv, January 22, 2025. https://arxiv.org/pdf/2501.12948.
Delaney, Oscar and Ashwin Acharya. “The Hidden AI Frontier.” AI Frontiers, August 28, 2025. https://ai-frontiers.org/articles/the-hidden-ai-frontier.
Department for Science, Innovation and Technology. “AI Safety Summit 2023: The Bletchley Declaration.” Gov. UK, November 2, 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.
Department for Science, Innovation and Technology. “International AI Safety Report 2025.” Gov. UK, January 29, 2025. https://www.gov.uk/government/publications/international-ai-safety-report-2025.
Department for Science, Innovation and Technology. “Frontier AI Safety Commitments, AI Seoul Summit 2024.” Gov. UK, February 7, 2025. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.
Drew, Rocket. “Will Chains of Thought Stay Readable for Long?” The Information, June 26, 2025. https://www.theinformation.com/articles/will-chains-thought-stay-readable-long.
Emmons, Scott, Erik Jenner, David K. Elson, et al. “When Chain of Thought Is Necessary, Language Models Struggle to Evade Monitors.” arXiv, July 7, 2025. https://arxiv.org/abs/2507.05246.
Erdil, Ege. “How Has DeepSeek Improved the Transformer Architecture?” Epoch AI, January 17, 2025. https://epoch.ai/gradient-updates/how-has-deepseek-improved-the-transformer-architecture.
Frontier Model Forum. “Frontier Model Forum: Advancing Frontier AI Safety and Security.” Accessed September 19, 2025. https://www.frontiermodelforum.org/.
Google Deepmind. “Gemini Diffusion.” May 20, 2025. https://deepmind.google/models/gemini-diffusion/.
Google Deepmind. “SynthID.” Accessed September 22, 2025. https://deepmind.google/science/synthid/ai-generated-text/.
Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. "AI Control: Improving Safety Despite Intentional Subversion." arXiv, July 23, 2024. https://arxiv.org/abs/2312.06942.
Grey, Markov, Charbel-Raphaël Segerie, Jeanne Salle, and Charles Martinet. “AI Safety Atlas.” French Center for AI Safety, 2025. https://ai-safety-atlas.com/chapters/01/06/.
Hao, Shibo, Sainbayar Sukhbaatar, DiJia Su, et al. “Training Large Language Models to Reason in a Continuous Latent Space.” arXiv, December 11, 2024. https://arxiv.org/pdf/2412.06769.
Harack, Ben, Robert F. Trager, Anka Reuel, et al. “Verification for International AI Governance.” Oxford Martin AI Governance Initiative, July 3, 2025. https://aigi.ox.ac.uk/publications/verification-for-international-ai-governance/.
Heim, Lennart, Tim Fist, Janet Egan, et al. “Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation.” arXiv, March 13, 2024. https://arxiv.org/abs/2403.08501.
Hendrycks, Dan and Thomas Woodside. “A Bird’s Eye View of the ML Field.” Center for AI Safety, April 10, 2024. https://safe.ai/blog/a-birds-eye-view-of-the-ml-field.

Ho, Lewis, Celine Smith, Claudia van der Salm, Joslyn Barnhart, and Rohin Shah. “Frontier Safety Framework.” Google Deepmind, February, 2025. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0.pdf.

Institute for Progress. “The Launch Sequence.” August 11, 2025. https://ifp.org/the-launch-sequence/.

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. “AI 2027.” AI Futures Project, April 3, 2025. https://ai-2027.com/ai-2027.pdf.

Korbak, Tomek, Mikita Balesni, Elizabeth Barnes, et al. “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.” arXiv, July 15, 2025. https://arxiv.org/abs/2507.11473.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. "Large Language Models are Zero-Shot Reasoners." 36th Conference on Neural Information Processing Systems (NeurIPS), 2022. https://arxiv.org/abs/2205.11916.

Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, et al. “Measuring Faithfulness in Chain-of-Thought Reasoning.” arXiv, July 17, 2023. https://arxiv.org/abs/2307.13702.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, et al. “The AI Index 2025 Annual Report.” Institute for Human-Centered AI, April 2025. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

METR. “Frontier AI Safety Policies.” February 17, 2025. https://metr.org/faisc.

Needham, Joe, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. “Large Language Models Often Know When They Are Being Evaluated.” arXiv, May 28, 2025. https://arxiv.org/abs/2505.23836.

OpenAI. “Preparedness Framework.” April 25, 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Schoen, Bronson, Evgenia Nitishinskaya, Mikita Balesni, et al. “Stress Testing Deliberative Alignment for Anti-Scheming Training.” Apollo Research, September 15, 2025. https://static1.squarespace.com/static/6883977a51f5d503d441fd68/t/68c9a63b9c1f2f236c7d97f6/1758045901755/stress_testing_antischeming.pdf.

Srinivasan, Siddarth. “Detecting AI Fingerprints: A Guide to Watermarking and Beyond.” The Brookings Institution, January 4, 2024. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

Studman, Anna, Mavis Machirori, Hannah Claus, and Imogen Parker. “Buying AI.” Ada Lovelace Institute, October 1, 2024. https://www.adalovelaceinstitute.org/report/buying-ai-procurement/.

The White House. “Winning the Race: America’s Action Plan.” July 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

Trask, Andrew, Aziz Berkay Yesilyurt, Bennett Farkas, et al. “Secure Enclaves for AI Evaluation.” OpenMined, January 2025. https://openmined.org/blog/secure-enclaves-for-ai-evaluation/.

Wang, Guan, Jin Li, Yuhao Sun, et al. “Hierarchical Reasoning Model.” arXiv, August 4, 2025. https://arxiv.org/pdf/2506.21734.