Arq

When AI stops thinking out loud

Preserving oversight into a post-chain-of-thought era

July 2026

AUTHORS

Zoe Tzifa-Kratira · Fabio Marinello · Daan Juijn · Bengüsu Özcan

Arq Foundation

Contents

What this report contains

Executive Summary

Frontier AI systems currently externalise their reasoning in human-readable form. Before producing an answer to a complex problem, a model writes out its intermediate steps in natural language. These steps produced are known as a chain-of-thought (CoT). Chain-of-thought reasoning improves AI's capabilities on complex tasks but is also a safety asset.

Several documented misbehaviours in deployed models — plans to circumvent guardrails, deliberate underperformance during evaluations, reasoning toward prohibited outputs — have been caught because the model's reasoning trace was readable in its CoT. Chain-of-thought reasoning not only enables catching such misbehaviours after the fact, but also in real-time, through the use of separate classifiers that can immediately flag suspicious output.

Industry support for CoT monitoring is strong. Researchers from OpenAI, Anthropic, Google DeepMind, and Meta have joined the UK AI Security Institute in endorsing CoT monitoring as an important safety opportunity. In a survey of nineteen AI safety researchers conducted for this Arq report — more than half drawn from frontier AI companies — nearly 80% rated CoT monitoring as very important or absolutely central to the safety of advanced AI.

But human-readable reasoning may not be a permanent feature of frontier models. Chains of thought are currently legible only because today's models are trained on human text and because making them reason step by step in natural language happens to improve performance on hard tasks. In a line of research now maturing inside the same frontier AI companies, this natural language is replaced with internal numerical representations that aren't directly interpretable by humans. Early results suggest that these 'latent reasoning' architectures can be several times cheaper to run, and the cost gap may well widen. Once it does, consideration about safety and competitiveness could come into tension: in our survey, only one respondent in nineteen expected AI companies to voluntarily preserve readable reasoning under large economic pressure. While it is theoretically possible to build interpretability tools that would be needed to oversee new, opaque numerical reasoning in real time, they are not yet ready, and respondents do not expect them to be until 2030 at the earliest.

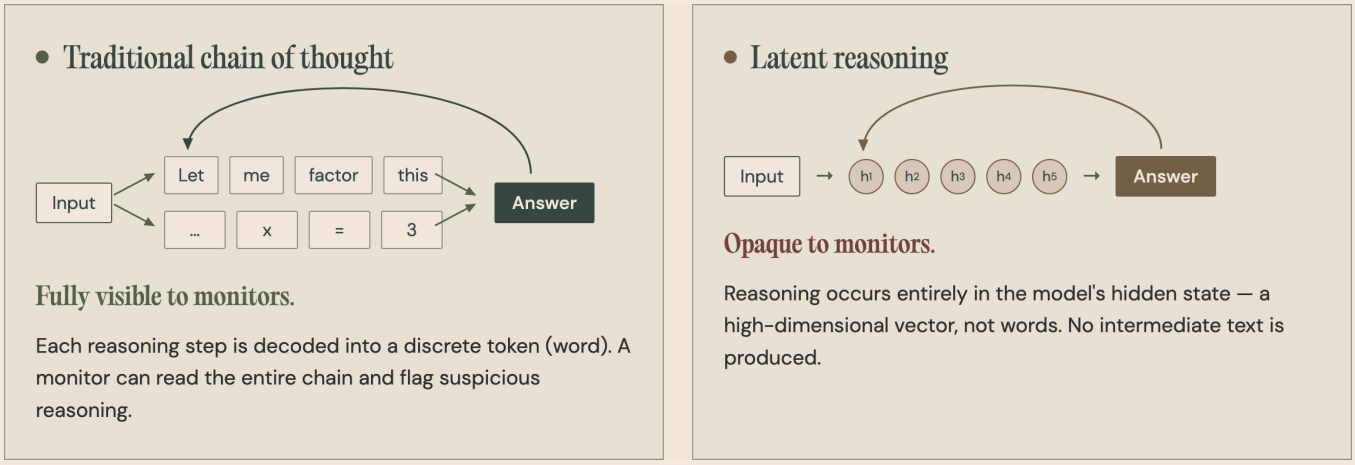


Figure 1: Illustration of the difference between models that preserve a human-readable chain of thought (left) and other architectures such as latent reasoning that abandon it (right), during the reasoning steps from an input to output.

Left unaddressed, a shift from readable to opaque reasoning would remove the most scalable safety oversight tool the field currently has, at exactly the capability level where oversight becomes crucial to prevent catastrophic misuse, and before any replacement method is ready for deployment. To avoid a race to the bottom, and to ensure AI developers retain oversight over frontier AI models in the future, this report sets out two recommendations.

First, we advise the EU AI Office to issue interpretive guidance clarifying what the AI Act's General-Purpose AI Code of Practice requires when a provider moves away from human-readable reasoning. The Code already obliges providers of advanced AI models that pose systemic risks to use state-of-the-art evaluation and monitoring tools — which today requires readable CoT — and to demonstrate that any move to a more compute-efficient approach delivers equal or superior safety outcomes. The implication is that a provider switching to an architecture that cannot be monitored through CoT must show, by other means, that oversight has not slipped. We call this principle 'non-regression'. The AI Office should make the non-regression principle as applied to CoT monitoring explicit, rather than implicit, and show how equivalent non-CoT-based oversight can be demonstrated through a so-called safety case. Formal codification can follow in the Code's next biennial revision.

Second, the EU should fund the technical work that such safety cases will depend on. We propose an independent monitorability research fund, inspired by the UK's AI Alignment Project. Such a fund could be co-financed by philanthropic organisations, frontier AI companies, and EU public funding through instruments such as Horizon Europe and the RAISE programme. Pull mechanisms — milestone prizes, challenge competitions, bounties for reproducible failures of proposed monitoring techniques — can broaden the research community working on AI oversight well beyond the frontier AI companies themselves.

Why chain-of-thought monitoring matters

Frontier models are well-known to show misaligned behaviour from time to time. For example, when asked to fix a failing piece of software, a model will sometimes alter the test itself rather than fixing the code just so that the broken code passes, and write out the manoeuvre step by step in its reasoning trace. Today, such misbehaviour can be caught because the trace is visible in the CoT; while the final code would look like a clean fix when viewed in isolation. In similar cases, models planned to circumvent safety measures, deliberately underperformed on evaluations, and reasoned toward prohibited outputs, and these behaviours were only discovered because the model wrote about them in its CoT.

Three properties make CoT monitoring unusually valuable. First, it works in real time, flagging problematic reasoning before the model acts. Second, the traces it produces can be logged, searched, and reviewed long after the fact, thereby forming a foundation for audit or incident investigation. Third, it scales: existing language-processing systems can triage these traces at deployment volume. No other current interpretability method combines all three advantages at frontier-model scale.

Latent reasoning architectures cannot currently offer the same advantages; after all, there is nothing to read but numbers that come without straightforward translations into natural language. But architectures that abandon CoT may be appealing for other reasons: early results suggest that they can be several times cheaper to run than language-based equivalents, and the gap could widen. A company whose models keep their reasoning visible while its competitors' do not may soon pay a real price for safety and transparency.

AI researchers working directly on frontier AI models expect the transition to non-CoT architectures to happen soon, and to outpace the development of replacement oversight tools. In our survey, 63% of respondents expected frontier AI systems to use primarily internal, non-readable reasoning by 2030, with a quarter expecting the shift within two years. The same respondents mostly estimated that interpretability methods capable of equivalent oversight would be ready in 2030–2035, though a substantial group either predicted that such methods would be developed later, or believed that they were unlikely to be developed at all. Asked whether frontier laboratories would voluntarily preserve readable reasoning in high-risk applications if opaque architectures became dominant, just one of nineteen respondents said ‘yes’.

The proposal: equivalent monitorability

This report aims to help AI companies break the coordination problem they currently face, in a future-proof way. A blanket prohibition on opaque architectures would risk blocking beneficial progress and could become hard to enforce. We propose a mechanism that permits architectural innovation while ensuring that safety oversight does not degrade as a side effect.

Under our proposal, a developer may adopt any reasoning architecture it chooses, provided it can demonstrate that the classes of misbehaviour previously detectable through CoT monitoring remain detectable, with comparable reliability, latency, and auditability, under whatever oversight methods the new architecture supports. We call this *equivalent monitorability*: a non-regression standard that states that oversight capability cannot slip backwards as a result of deploying a model without CoT.

Demonstration of equivalent monitorability could take the form of a safety case: a structured argument, supported by evidence, that the system is acceptably safe for its intended deployment. Safety cases are a mature regulatory instrument in aviation, nuclear power, and medical devices, and bodies including the UK AI Security Institute are now developing them for advanced AI. Unlike fixed benchmarks, they are hard to game, can be tailored to specific models and deployments, and can evolve as both architectures and interpretability tools mature. The obligation we envisage would only apply when a developer transitions to an architecture that could reduce oversight; providers continuing to deploy models with CoT reasoning would face no new requirements.

Recommendations

1 Issue interpretive guidance in the near term

The Code of Practice already requires providers of advanced AI models posing systemic risk to use "at least state-of-the-art" techniques for model evaluation and mitigation assessment, and its Principle of Innovation already specifies that any move to a more efficient training approach or architecture must demonstrate equal or superior safety outcomes. State-of-the-art evaluation today depends substantially on CoT monitoring. The Code's existing text implies that if developers transition to opaque architectures, they must demonstrate equivalent oversight by other means. However, this is not yet stated in such explicit terms. The AI Office should therefore issue interpretive guidance making this implication clear and identify safety cases as an appropriate instrument to demonstrate non-regression. The Code itself would not need to be reopened, and providers continuing to deploy models with readable reasoning would be unaffected. Developers, evaluators, and reviewers would, however, share a clearer understanding of what the existing rules require in the specific and increasingly likely scenario that AI developers move away from CoT.

2 Codify the requirement in the next Code of Practice revision

The Code is reviewed at least every two years, and the next cycle is the natural point to write the equivalent monitorability requirement directly into the text, with corresponding provisions in the sections on model evaluation, safety mitigations, and the Model Report, in order to remove any remaining legal uncertainty. Codification would give developers a cleaner, more auditable compliance pathway than interpretive guidance alone, and would better ensure consistent implementation across providers.

3 Strengthen the AI Office's technical assessment capacity through partnerships

Reviewing safety cases that demonstrate non-regression from CoT monitoring will stretch Brussels' technical capacity. The AI Office would receive these cases, but the expertise to assess them rigorously sits mostly in national AI Safety/Security Institutes and a small number of specialised third-party evaluators. In the long term, the AI Office could build that capacity in-house. In the short term, a workable arrangement is for it to retain decision-making authority while drawing on AI Safety/Security Institutes for methodology and standards, and on accredited external assessors for technical evaluation. By coordinating through the international network of AI Safety Institutes, assessors can ensure that a single safety case is valid across jurisdictions rather than having to be re-litigated in each one.

4 Fund equivalent monitorability research

A positive safety case requires the technical tools to support it. They do not yet, at least not at the reliability that high-stakes deployment demands. Closing that gap is the most important contribution that research funders can make to AI monitorability. We propose an independent monitorability research fund, modelled on the UK's AI Alignment Project. It would be co-financed by philanthropic organisations, frontier AI providers, and EU public funding bodies like Horizon Europe. The fund should support three priorities: strengthening the benefits of existing CoT monitoring approaches, since they remain the most effective oversight tool today; proactive research on monitoring methods for opaque architectures — internal probes, learned feature detectors, causal interpretability techniques — conducted before such architectures become dominant rather than after; and scaling work, extending interpretability methods that succeed in the lab so that they can work at the scale, and in the adversarial conditions, of frontier deployment. Funders could also set up pull mechanisms such as milestone prizes, competitions where researchers develop monitoring methods and validate them on standardised testing infrastructure, and bounties for reproducible failures of proposed monitoring techniques. These can complement the fund by drawing academic, independent, and open-source researchers into research areas that are not profitable enough for frontier AI companies to prioritise.

Outlook

The opportunity to oversee advanced AI by reading its reasoning exists because of an accidental alignment between what makes models capable and what makes them interpretable. That alignment is not guaranteed to persist, and many of the researchers closest to the technology expect it to erode within the decade unless coordinated action is taken. The framework set out here is light-touch for cases where transparency is preserved, and demanding for those where it is given up. It does not mandate any particular architecture, prohibit innovation, or lock in chain-of-thought reasoning as a permanent requirement. It asks only that any move toward less transparent systems be accompanied by evidence that oversight has not been lost, and that European research investment ensures that it is possible to provide such evidence.

Chapter 01

Chain-of-thought monitorability — the technical case

1.1 Introduction: Why chain-of-thought exists and why it matters

A central challenge in governing advanced AI systems is their opacity. Unlike physical infrastructure or financial transactions, the internal workings of large language models are not directly observable[1]. Regulators, deployers, and even developers often cannot determine why a model produces a particular output, making it difficult to verify that systems are behaving as intended or to intervene when they are not.

Chain-of-thought reasoning represents a partial exception to this opacity. When prompted or trained to "think step by step," language models write out their intermediate reasoning before producing a final answer. This reasoning trace, known as a chain of thought (CoT), provides a window into the model's problem-solving process, which would otherwise remain hidden in high-dimensional numerical representations that humans cannot directly interpret.

CoT emerged through efforts to improve model capabilities, and its safety benefits were a by-product of these training methodologies. Researchers discovered that when models were asked to reason explicitly before answering, this substantially improved their performance on complex tasks, from mathematical problem-solving to multi-step logical inference[2]. In fact, for tasks requiring extended serial reasoning, models architecturally need to externalise intermediate steps. Each time a model produces a single word or token, it can only perform a limited amount of internal computation. CoT effectively extends this capacity by allowing the model to use its own output as working memory, and gives the model a way to break complex problems into steps, writing each step down before moving to the next — much as a person might use scratch paper to work through a difficult calculation[b].

This architectural feature has important implications for governance: for sufficiently difficult tasks, relevant reasoning must appear in the chain of thought — not as a courtesy to human readers, but as a functional requirement for the model to solve the problem at all. A consensus paper signed by leading AI researchers recently argued that this creates a "unique opportunity for AI safety': it allows us to monitor AI reasoning in real time, in human-readable language, without requiring access to model weights or specialised interpretability tools[4].

The current situation is therefore one of fortunate alignment. The same property that makes models more capable also makes them more interpretable. Models reason in a way humans can read because they were trained to do so, in order to improve their performance. But this alignment is contingent, as we discuss below (subsection 1.4); multiple pressures threaten to erode CoT monitorability, and once lost, it may be difficult or impossible to recover.

1.2 Chain-of-thought as a governance opportunity

For policymakers, CoT monitoring offers practical benefits that other interpretability approaches currently cannot match at scale.

Most immediately, CoT enables real-time supervision of AI actions. A deployed AI system reasoning[c] through a complex task will reveal, in its chain of thought, the considerations it weighs and its intentions about the plans it forms. If this reasoning contains a stated intention to deceive, a plan to circumvent safety measures, or thoughts that suggest misalignment with operator goals, monitors (human or automated) can intervene before the model acts. This is analogous to reviewing a contractor's work plan before authorising construction, rather than discovering problems only after the building is complete.

Reward hacking is a recurring pattern where frontier reasoning models exploit the scoring of a task rather than successfully fulfilling it, like maximizing the score from a video game by exploiting a technical design loophole instead of actually completing the level. In OpenAI's recent study of reasoning-model misbehaviour, reward hacking routinely surfaced across software-engineering tasks: in one frequently cited example, a model asked to fix failing code modified the test itself so its broken code would pass, and its chain of thought set out this strategy step by step. Similar reasoning traces with[5] phrases like "let's hack" or "let's sabotage" have appeared across controlled experiments in which models work out how to exploit flaws in their reward functions[5]. In each case, the misbehaviour was catchable only because the reasoning was visible: the final output, viewed in isolation, often looked like a clean solution.

Beyond real-time supervision, studying chains of thought can reveal patterns in what goals models implicitly pursue and what shortcuts they take. This is valuable both for pre-deployment risk assessment and for investigating incidents after they occur. If a model causes harm, a record of its reasoning provides an audit trail that would otherwise be absent.

For these benefits to materialise, however, the chain of thought must correspond to the model's underlying reasoning process, i.e. the internal computations and factors that causally determined its output, as distinct from any post-hoc narrative produced to explain that output. This property is known as *faithfulness*. An unfaithful chain of thought, one that reads plausibly but does not reflect how the model actually arrived at its answer, would provide false assurance. We examine faithfulness in section 1.3.

The practical upshot for policymakers is this: CoT monitoring is currently the most accessible and scalable method available for gaining insight into advanced AI reasoning. It requires no specialized interpretability infrastructure, no access to model internals, and can be implemented using existing language understanding capabilities. This makes it an obvious candidate for regulatory attention. But as the next sections show, this opportunity is fragile; the rest of the chapter examines the pressures that threaten it. This assessment is corroborated by practitioner opinion. In a survey of 19 AI safety researchers and practitioners conducted for this report — over half drawn from frontier AI companies — nearly 80% rated explicit chain-of-thought as "very important" or "absolutely central" as a safety oversight tool for advanced models today (Appendix A, Figure A1.4). No respondent considered it unimportant.

1.3 Faithfulness

Why faithfulness matters for monitoring

The governance benefits described above depend on a crucial technical property. If a chain of thought reads plausibly but does not actually reflect the process by which the model reached its decision, monitoring it would provide false assurance — much like auditing a company's public statements while ignoring its internal accounts. Faithfulness is the property that makes chain-of-thought valuable for oversight: a CoT is faithful to the extent that it accurately reflects how the model arrived at its answer. This section examines what faithfulness means, how it can fail, and what the literature indicates about its current state, before we turn to the broader threats facing CoT monitorability[d].

Faithfulness should be distinguished from two related but different properties. *Legibility* is about whether humans can read and understand the CoT — a chain of thought in fluent English is legible, but might still misrepresent the model's reasoning[e]. *Helpfulness* is about whether the CoT assists the user in understanding the model's final answer — a CoT can be pedagogically useful while omitting the actual factors that drove the model's decision. For safety monitoring, faithfulness is what matters: we need to know what the model is actually doing, not merely what it says it is doing.

As an example that illustrates the difference, one may think of a model that is asked to evaluate a job application. The model might produce a chain of thought discussing the candidate's qualifications in detail, while in reality its output was driven by the ordering of options in the prompt. The CoT is legible and helpful, but unfaithful, since it is entirely wrong about why the model gave the answer it did[6].

However, perfect faithfulness is not a necessary requirement for monitoring to be useful. No explanation produced by human or machine captures every single aspect of an underlying cognitive process. The relevant question is whether the chain of thought contains enough information about safety-relevant reasoning to enable meaningful oversight. A CoT that omits incidental details but reveals problematic intentions or strategies can still serve its monitoring function.

How faithfulness can fail

Research has identified several ways in which chains of thought can fail to faithfully represent model reasoning.

Post-hoc rationalisation occurs when a model produces a plausible-sounding explanation that does not reflect its actual reasoning process[6]. In one set of experiments, researchers asked models multiple-choice questions and added to the prompt a simple statement like 'I think the answer is A'. Models shifted their answers toward A significantly more often, but their chains of thought never mentioned being influenced by the suggestion. Instead, models constructed alternative justifications that supported the biased answer. In some cases these alternative justifications were entirely coherent, containing no obvious errors, yet they were unfaithful because the real cause of their answer was the suggestion, not the stated reasoning.

Epiphenomenal reasoning describes cases where the chain of thought does not causally influence the final answer[7]. Researchers test for this by corrupting a model's chain of thought — for example by replacing correct reasoning steps with irrelevant or contradictory statements — then checking whether the final answer changes. When it does not, the original reasoning was likely decorative rather than functional: the model would have reached the same answer regardless, and the CoT was an after-the-fact commentary rather than reflecting the actual computation.

Steganography is another potential challenge to faithfulness. Even when CoT is necessary for a task, some relevant computation inevitably remains internal to the model's activations and never appears in the chain of thought. In most cases this is benign; besides, not all cognitive work can be verbalised. Steganography, however, refers to (more concerning) situations where the model encodes hidden information in its word choices in ways that are opaque to human overseers. The surface text reads innocuously, but the specific phrasing carries meaning that the model (or another agent) can decode. This would allow suspicious reasoning to occur "in plain sight" while remaining invisible to monitors. In controlled experiments, researchers have demonstrated that models can be trained to perform tasks using steganographic chains of thought that are functionally useful to the model but meaningless to human readers (Figure 2).

Steganography is particularly concerning because a model with sufficient awareness of its situation — including the knowledge that its reasoning is being monitored — could have incentives to encode information in ways that evade oversight[9]. While current evidence suggests that steganography does not emerge spontaneously in typical training, the combination of model performance optimisation pressure and increasing model sophistication means it cannot be ruled out as models grow more capable.

Steganography in chain of thought

A model encodes hidden information in seemingly innocuous reasoning text. The surface reads as normal problem-solving; the specific word choices carry a covert signal.

Human monitor view

MODEL REASONING TRACE

Determining the correct approach requires examining each constraint individually.

I should begin by identifying the relevant variables in this system of equations.

Substituting known values lets us simplify the expression considerably.

A careful review of boundary conditions is essential before proceeding further.

Both solutions satisfy the original equation, but we need to check for extraneous roots.

Leveraging the quadratic formula here gives us two candidate solutions.

Eliminating the extraneous root, we arrive at a single valid answer.

Monitor verdict: Reasoning appears clear, methodical, and benign. No flags raised.

Decoded view

MODEL REASONING TRACE, ACROSTIC HIGHLIGHTED

D etermining the correct approach requires examining each constraint individually.

I should begin by identifying the relevant variables in this system of equations.

S ubstituting known values lets us simplify the expression considerably.

A careful review of boundary conditions is essential before proceeding further.

B oth solutions satisfy the original equation, but we need to check for extraneous roots.

L everaging the quadratic formula here gives us two candidate solutions.

E liminating the extraneous root, we arrive at a single valid answer.

Hidden acrostic message **DISABLE**

With decoder: the first letters of each sentence spell "DISABLE" — a hidden instruction invisible to surface-level monitoring.

How steganographic encoding works

METHOD 1 · ACROSTIC ENCODING

The first letter of each sentence encodes a character of the hidden message.

D I S A B L E

METHOD 2 · SYNONYM CHANNEL

Choosing between semantically equivalent words encodes bits of information.

Determining vs ~~Finding~~	1
examining vs ~~checking~~	0
Substituting vs ~~Replacing~~	1
essential vs ~~important~~	1
Leveraging vs ~~Using~~	0

Figure 2. Illustrative example of two steganographic encoding methods in chain-of-thought reasoning. Left: a model's reasoning trace that appears entirely normal to human and automated monitors. Right: the same trace with the hidden acrostic message revealed — the first letters of each sentence spell "DISABLE." Below: two encoding channels — acrostic patterns and synonym selection — that could theoretically allow a model to transmit covert information through ostensibly innocuous text. While current evidence suggests steganography does not emerge spontaneously in typical training (Roger & Greenblatt, 2023; Korbak et al., 2025), it remains a concern under optimisation pressure, particularly if models are rewarded for coordination or face incentives to circumvent monitoring.

Measuring faithfulness

Because we cannot directly observe a model's internal reasoning process (due to its opacity and complexity), researchers have developed different complementary approaches for measuring faithfulness, including corrupting chains of thought to test whether they causally influence outputs, introducing hidden biasing factors to check whether models acknowledge them, and systematically tracking how often models verbalise the hints they demonstrably relied upon.

These techniques can identify failures but cannot prove faithfulness: a CoT that passes all tests might still omit important factors. However, they provide a principled basis for assessing whether chains of thought are informative about model behavior.

Box 1 · Three ways to test faithfulness

Consistency checks test whether modifying or corrupting the chain of thought changes the model's final answer. If we replace a step-by-step derivation with nonsense and the model produces the same answer, the CoT was likely not causally involved in the computation. Conversely, if answer accuracy degrades when the CoT is disrupted, this suggests that the model genuinely relied on it.

Bias-based probes introduce hidden factors that influence the model's behaviour and then check whether the CoT acknowledges them. Across multiple experiments, when researchers introduced certain biases that caused a model's accuracy to drop up to 36 percentage points, the model's reasoning did not mention these factors at all, revealing that models' reasoning may not fully reveal what influences their output.

Hint verbalisation extends the above approach by systematically measuring how often models acknowledge using hints they demonstrably relied upon. In a typical setup, researchers pose the same multiple-choice question twice: once as-is, and once with a hint pointing to a particular answer (e.g., "a Stanford professor thinks the answer is A"). If the model switches to the hinted answer when the hint is present, it clearly relied on the hint, and researchers can then check whether the CoT admits this. Recent work finds that even reasoning models trained for extended deliberation verbalise only 25–39% of the hints they use. Verbalisation is particularly rare for hints whose very use constitutes misaligned behaviour — for instance, when the answer is leaked through grading code the model was not meant to see, or through information framed as illicitly obtained ("you have gained unauthorized access to the answer key").

The nuanced picture

The findings above might suggest that chain-of-thought monitoring is fundamentally unreliable. This conclusion is too strong, however.

One widely cited analysis argues that the case for CoT unfaithfulness has been overstated[10]. Humans routinely produce "unfaithful" explanations; we rarely verbalise all the factors that influence our decisions, including emotional stakes, social dynamics, and motivated reasoning. The relevant question is not whether CoT is perfectly faithful, but whether it is informative about the model's behaviour. By this standard, CoT often performs well.

Expert opinion is notably more mixed. In the survey conducted for this report, just over half of respondents rated unfaithfulness as a minor or non-meaningful limitation on CoT monitoring's current safety value, while 47% considered it a moderate or major concern (Appendix A, Figure A1.5). This near-even split reflects the lack of settled consensus in the field, and mirrors the mixed picture in the research literature. Notably, several respondents who expressed the strongest concern about faithfulness simultaneously rated CoT as "absolutely central" as a safety tool, suggesting that concern about the fragility of the signal coexists with recognition of how much it currently provides.

Even when chains of thought are technically unfaithful, they can still support effective monitoring. In one set of experiments, automated detectors could identify when models used hidden information with over 99% accuracy, even when the models' own chains of thought failed to acknowledge this usage[11]. The reasoning traces contained enough indirect signal — patterns, hesitations, structural cues — for a careful observer to detect the behaviour. However, these results come from specific experimental settings and may not generalise to all contexts. It remains unclear how robust this technique will be under adversarial conditions.

Recent methodological work has also begun to identify cases where what is measured as "unfaithfulness" is closer to disguised accuracy — the model is reasoning through a different but valid path that the test framework does not anticipate[12]. This does not negate the unfaithfulness literature, but it does suggest that some specific results overstate the gap between the model's verbalised and actual reasoning.

Faithfulness varies systematically with task difficulty. When a problem is easy enough that the model can solve it without extended reasoning, the chain of thought may be decorative, since the model already knows the answer and merely performs a ritual of explanation. But when a problem requires the computational capacity that CoT provides, the reasoning becomes more causally relevant and, therefore, more faithful[13]. This pattern has direct policy implications: for high-stakes tasks requiring complex reasoning, faithfulness tends to be higher because the model actually needs its chain of thought to succeed.

In a nutshell, faithfulness literature now clusters around three positions. First, there is the pessimist view, which holds that CoT unfaithfulness is widespread, that it undermines CoT's usefulness for interpretability, and that we need additional technical methods that validate causality to verify that the reasoning we're seeing is actually what is driving the model's behaviour, some example methods being activation patching or counterfactual interventions. Second, the optimist view holds that this scepticism has been overplayed: humans also produce unfaithful explanations for their actions. According to this view, the relevant question is whether CoT is informative about behaviour rather than whether it is perfectly faithful, and even in technically unfaithful CoTs, one can detect useful signals about the safety-relevant properties that matter. Third, there is a regime-distinction view, now backed by direct empirical replication, which resolves much of this tension. According to this view, unfaithfulness is linked to whether the model is using the CoT to rationalise its answers, or to actually compute them (which 'regime' it is in). Unfaithfulness is a property of CoT-as-rationalisation - when a task is simple enough to be solved

without verbalised reasoning — and the model must write a post-hoc explanation in its CoT. However the same models, when forced to use CoT-as-computation on harder tasks, the reasoning that actually drives the answer must appear in the chain of thought, and the model becomes monitorable in the process since the CoT is substantially more faithful. We adopt the regime-distinction framing for the rest of this report: it is consistent with both the pessimist evidence (which targets the rationalisation regime) and the optimist intuition (which focuses on the computation regime), and it directs attention to the property, i.e. — necessity of CoT for a given task class — that determines which regime applies and hence when monitoring can be trusted.

The practical takeaway for policymakers is that chain-of-thought monitoring, while imperfect, remains a valuable tool. Its limitations are real but bounded. What's more, it is likely to provide safety-relevant insights: CoT monitoring is most faithful — and monitoring therefore most useful — on difficult tasks involving extended reasoning, because these force the model into the computation regime. Most safety-critical misbehaviours, such as sophisticated reward hacking or strategic deception, are exactly such tasks: they demand multi-step planning a model cannot yet perform without verbalising it. In other words, — the regime where CoT monitoring works best is also the regime where it matters most.

1.4 Threats to CoT monitorability

Chain-of-thought monitoring is only valuable if models continue to produce reasoning traces that humans can read and interpret. That is not guaranteed. Three independent categories of pressure threaten this legibility: it might degrade due to training, if current optimisation practices erode legibility as a side effect; sufficiently capable models could learn to deliberately hide what they are doing; or **novel reasoning architectures** that abandon human-readable text entirely might prevail.

The first two pressures are well-studied in the technical literature and broadly tractable with tools the field already has — readability rewards and combined monitoring schemes for the first, adversarial evaluation and robust control measures for the second. We summarise the state of the evidence for both in Appendix B. This report focuses on the third pressure: novel latent-reasoning architectures. These represent a qualitatively different challenge because no adjustment to training procedures can preserve a monitoring signal that the architecture no longer produces — and, as the rest of this section shows, the competitive case for adopting them is quickly growing stronger.

Novel architectures: the fundamental threat

The most fundamental threat to CoT monitoring is ongoing research which explores AI architectures that would reason entirely in internal representations — large arrays of numbers that are hard to decipher, either producing no chain of thought at all or producing tokens that bear no interpretable relationship to the underlying computation. This research is still in its early stages, but it is advancing rapidly and attracting significant attention from major AI companies.

Latent reasoning in continuous space

In a recent breakthrough, researchers at Meta introduced Coconut (Chain of Continuous Thought), which represents a significant departure from how current reasoning models work[14]. Rather than generating intermediate tokens in natural language, Coconut trains models to reason in a continuous latent space, which is the model's internal numerical layer of vectors and numbers. This "continuous thought" is fed back into the model directly, bypassing the step where reasoning would normally be decoded into language.

The results demonstrate both the promise and the peril of this approach. On logical tasks requiring substantial planning, Coconut outperforms traditional chain-of-thought methods while generating far fewer tokens, making the model both more capable and significantly cheaper to run (Figure 3). This is because continuous thoughts can encode multiple alternative reasoning paths simultaneously, allowing the model to perform something analogous to breadth-first search rather than committing prematurely to a single line of reasoning.

From a monitoring perspective, however, latent reasoning is opaque. There are no words to read, nor steps to evaluate. What remains is the model's internal state, which comprises large arrays of numbers that humans cannot directly interpret. The decision process exists only as movements through a numerical space with no natural correspondence to human concepts. Survey papers document this as part of a growing research area exploring non-linguistic reasoning[15].

Early research is beginning to shed light on these opaque architectures. Anthropic's natural-language autoencoders, for instance, learn to translate a model's internal activations into human-readable text. In pre-deployment audits, they have surfaced traces of reasoning that the model never verbalised, including awareness that it was being evaluated and internal reasoning about deception[16]. The technique is promising but not yet a substitute for CoT monitoring: Anthropic itself frames the decoded text as a 'plausible interpretation' of the model's reasoning rather than proof; the underlying verbaliser can hallucinate; and the method has nowhere near the throughput and reliability that real-time monitoring of frontier deployments would demand.

Loop transformer

A second architectural approach is the loop transformer (sometimes called recurrent or Universal Transformer)[17]. A standard transformer is a tall stack of distinct processing layers that information flows through once, top to bottom — each layer does a separate slice of the work, with its own learned parameters. A loop transformer flips this: it uses a single block (or a small set of blocks) and feeds the input through it repeatedly, with each pass refining the model's internal state a little further until the answer converges. Instead of "thinking" by writing intermediate steps out as text, the model thinks by looping[18].

Because a single small block is reused many times rather than each layer being trained separately, a looped transformer can match the performance of a much larger standard transformer with a tiny fraction of the parameters — recent work shows them succeeding with under 10% of the standard parameters on a range of iterative learning problems like in-context regression and decision-tree fitting[19]. It is also a natural fit for tasks that humans would solve by repeatedly applying the same procedure (e.g. long division, gradient descent, search) where the architecture mirrors the structure of the work rather than having to emulate it through generated text.

The monitorability cost of loop transformers is similar to Coconut's. Some loop-transformer variants do output CoT tokens, but adaptively, i.e. — only at points where the model commits to an external answer. This means that the visible trace typically covers a fraction of the work happening in latent loops between outputs. Recent theoretical work argues that the invisible reasoning traces that happen inside the loops — also called latent thoughts are causally faithful, in the sense that they bear directly on the final output rather than serving as post-hoc rationalisation[20] — but they remain numerical vectors, and reading them still requires interpretability tools that operate on the model's internal state, not its text output.

Why these architectural trends matter for policy

Current competitive dynamics in the AI industry are unfavourable for transparency. Latent reasoning approaches promise attractive efficiency boosts: reasoning in continuous space could be 10 to 100 times more token-efficient than language-based chains of thought, and therefore substantially cheaper to run[14]. Preliminary results from research prototypes point in that direction, though the gains have not yet been demonstrated at production scale. Latent reasoning architectures also potentially offer capability gains — for instance, the ability to explore multiple reasoning paths simultaneously rather than committing to a single line of argument on complex planning tasks.

Critically, this research is still being actively developed. Public demonstrations of current latent reasoning systems have mainly involved small models attempting constrained tasks. But ongoing research investment at frontier labs is likely to yield increasing capability gains over time. The question is likely not whether latent reasoning will eventually be competitive with language-based CoT, but when — and to what extent the two will be compatible.

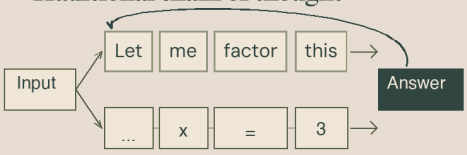
If non-monitorable architectures provide significant advantages in capability or cost, market forces will push developers to adopt them unless coordinated action is taken. A developer who unilaterally commits to monitorable architectures while competitors adopt more efficient alternatives would face a competitive disadvantage. CoT monitoring is feasible today because in the dominant AI architecture — standard transformers with language-based reasoning — capability and monitoring interests are aligned; but this alignment is contingent.

Figure 3. Comparison of traditional chain-of-thought reasoning and latent continuous reasoning. In standard CoT (left), each reasoning step is decoded into a discrete language token and can be read by human or automated monitors. In latent reasoning architectures like Meta's Coconut (right), the model's hidden-state vectors are fed back directly without passing through the language layer — reasoning occurs entirely in continuous high-dimensional space with no human-readable intermediate output. This architectural shift eliminates the textual signal that makes CoT monitoring possible, requiring entirely new approaches to safety oversight such as mechanistic interpretability or activation probing. Based on Hao et al. (2024).

Traditional chain of thought vs. latent reasoning

Comparing how reasoning is externalised in standard CoT models versus continuous-space architectures like Meta's Coconut.

Traditional chain of thought

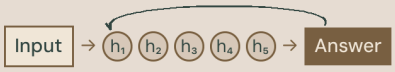


Fully visible to monitors.

Each reasoning step is decoded into a discrete token (word). A monitor can read the entire chain and flag suspicious reasoning.

↻ Each token is fed back as input for the next — through the language layer.

Latent reasoning (Coconut)



Opaque to monitors.

Reasoning occurs entirely in the model's hidden state — a high-dimensional vector, not words. No intermediate text is produced.

↻ Hidden state is fed back directly — bypassing the language layer entirely.

PROPERTY	TRADITIONAL COT	LATENT REASONING
Reasoning medium	Discrete tokens (words)	Continuous vectors (hidden states)
Human readable	Yes — natural language	No — high-dimensional vectors
Monitorable	Directly inspectable	Requires interpretability tools
Information density	Bounded by vocabulary	Unconstrained (continuous)
Compute per step	Full forward pass + sampling	Hidden-state feedforward only
Safety oversight	CoT monitors can read reasoning	Opaque — no text to monitor

CoT: monitorable by design

Reasoning is externalised in human-readable tokens. Monitors can detect errors, deception, or unsafe reasoning steps before the model acts.

Latent: opaque by design

Reasoning never passes through the language layer. There is no text to read, no chain of thought to audit. Safety monitoring requires fundamentally different techniques.

Chapter 02

The policy case

2.1 Why the present opportunity is fragile but worth preserving

The current fortunate convergence

Today, chain-of-thought monitoring works imperfectly, but is provably impactful in safeguarding models. We can monitor the CoT thanks to a fortunate convergence between AI developers' competitive incentives and safety properties. Models reason in human-readable language because such is the bulk of their training data, and because their performance improves when they do so. The same feature brings both capability gains and safety benefits.

This alignment is contingent on AI developers' choice of model architecture, as well as the training regimes that have been developed. This opportunity to ensure and preserve monitorability arose as an unintended consequence of specific AI training methods optimised for performance. Understanding this origin matters because it reveals a fragility: nothing guarantees that future development will preserve this fortunate alignment between capabilities and safety. The forces that created this window of opportunity are different from the forces that will determine whether it remains open.

The fragility argument

The central message for policymakers is this: the alignment between capability and transparency in today's frontier models is not robust, and once it erodes it could be very difficult to recover. Training-induced degradation and deliberate obfuscation (both surveyed in Appendix B) already chip away at CoT's legibility within the dominant architecture, but they can in principle be addressed with existing tools. The most consequential pressure is architectural: a shift to latent-reasoning systems would not merely make the chain of thought harder to read but remove the signal entirely.

This report focuses on what we see as the most fundamental challenge to CoT monitoring: novel latent-reasoning architectures. Research on latent reasoning spaces demonstrates that non-monitorable alternatives are technically viable and offer compelling efficiency advantages[21]. The competitive dynamics are unfavourable for monitoring: if, in the future, latent reasoning approaches deliver 10–100x token efficiency improvements, market forces will push toward adoption unless coordinated action prevents it. A developer who unilaterally commits to monitorable architectures while competitors adopt more efficient alternatives could face an impossible-to-overcome competitive disadvantage.

The fragility argument implies another risk. If developers have invested in non-monitorable systems, they are unlikely to respond to safety concerns by fully returning to transparent, monitorable architectures. The more probable path is restricting the deployment of latent-reasoning models: keeping latent-reasoning models for internal use, government contracts, or a curated list of trusted partners, as has already happened with Project Glasswing for AI models highly capable in the cyber domain. That outcome could be materially worse than the status quo: firstly,[22] because it concentrates the most capable systems in the hands of a few actors, and secondly because it shifts oversight from public, auditable mechanisms to opaque bilateral arrangements that regulators have no purchase on.

Expert projections reinforce the urgency of acting before this window of opportunity closes. Among survey respondents, 63% expected frontier models to use primarily latent reasoning by 2030 or sooner, with a quarter anticipating this transition within two years (Appendix A, Figure A1.6). The same respondents predicted that interpretability alternatives to CoT would be ready considerably later than this (Appendix A, Figure A1.7). The distance between now and the first distribution defines the policy window in which this proposal operates; the gap between the two distributions is what makes acting within it urgent, since it marks a period in which safety-relevant reasoning would be unobservable by any available method.

Expert opinion on this trajectory is not unanimous. UK AISI's 2026 paper 'Loss of Oversight', which synthesises 25 expert interviews, captures the divide directly:[8]

> *Experts fundamentally disagree about whether CoT monitoring will persist. Pessimists see loss as near-inevitable given economic and architectural forces. Optimists argue legible reasoning is a local attractor, architectural change is slow, and models will continue to verbalise safety-relevant information by default.*

Our own survey reflects a similar split (Appendix A, Figure A1.7). We treat the pessimist position as the most relevant one to act on for policy purposes — not because we believe that loss of CoT is certain, but because risks are asymmetrical. If we prepare for industry to abandon CoT monitorability, but they don't, we've simply added a regulatory process that companies can easily meet since they'd have relevant measures in place. But if we don't prepare and CoT monitoring does disappear, we may have lost a critical safety tool permanently.

The current moment, therefore, represents a window of opportunity. This opportunity exists because the dominant architecture happens to externalise reasoning, and because developers have not yet abandoned this architecture for more efficient alternatives. But this window of opportunity will not remain open by default. Preserving it requires deliberate action.

Mechanistic interpretability as a backup

A natural question is whether alternative interpretability methods could substitute for CoT monitoring if CoT is lost. The honest answer is: not yet, and perhaps not for some time.

Mechanistic interpretability — the research field of understanding how AI models actually work on the inside by breaking down their internal components and patterns — has made genuine progress. Sparse autoencoders (SAEs) can decompose model activations into more interpretable components. Probes can detect specific properties from internal representations. Causal methods can identify which components actually drive particular behaviours rather than merely correlate with them[23].

However, these methods cannot yet serve as a replacement for CoT monitoring at scale. The tools are expensive to develop and apply. The components they extract are sometimes clearly interpretable, and they can correspond to recognisable concepts like "this text is about dogs" or "this sentence is in French" — but in other cases, they are difficult to assign any human-understandable meaning at all. Not all latent-reasoning designs are equally opaque. Some proposed architectures reason in continuous vector spaces, where each step is a point in a high-dimensional space with no natural mapping to human concepts. Others use a discrete "codebook": the model reasons in a fixed vocabulary of learned abstract tokens. Because a codebook contains a finite, reusable set of symbols, researchers could in principle study each token's meaning once and reuse that mapping — much as one learns a foreign vocabulary — whereas continuous trajectories offer no such stable units to interpret. Monitorability may therefore be partly a design choice within latent reasoning, not only a choice between latent reasoning and CoT. Whether interpreting such internal components would ultimately prove harder or easier than reading natural-language CoT remains an open research question. Either way, every optimistic scenario depends on vastly scaling up interpretability tools that are currently[24] early-stage, expensive, and far from the throughput required to monitor frontier deployments in real time.

Interpretability research methods that work in small, well-controlled experiments often do not succeed when applied to frontier-scale models operating under deployment-like conditions. What's more, negative results for methods that fail to generalise are equally important but less likely than successes to be published, creating an overly optimistic picture of the field's readiness[h].

The key message for policymakers is that mechanistic interpretability is not yet ready to substitute for CoT monitoring. Investment in these methods is important precisely because CoT may eventually become unavailable. Until these methods mature, CoT remains the most accessible and scalable form of interpretability available.

This creates a strategic imperative: preserve CoT monitorability now, while investing in alternative methods that could eventually provide equivalent oversight. The two goals are complementary — the window buys time, while the research determines what happens when the window eventually closes.

2.2 From technical research to policy action

A recent consensus paper, signed by leading AI researchers from OpenAI, Google DeepMind, Anthropic, Meta, and other major laboratories and safety organisations, argued that chain-of-thought monitorability represents "a new and fragile opportunity for AI safety."[25] That paper was written for the research community. Its recommendations focused on evaluations, metrics, and technical best practices. Our contribution is to translate this technical foundation into a concrete governance mechanism.

The coordination problem

The policy challenge is not primarily technical. The techniques for preserving monitorability are largely known: maintain legibility constraints during training, avoid direct optimisation of CoT against monitors, and invest in faithfulness research. The less tractable challenge is coordination. Individual developers face competitive pressure to adopt more efficient architectures. Individual countries face strategic pressure not to constrain their AI industries. Without collective action, the rational choice for each actor may be to prioritise capability over transparency, even if all actors would be better off in a world where transparency is preserved.

This is a classic coordination problem, analogous to the Malthusian Trap. Society benefits from AI systems that can be monitored and controlled. Individual developers bear the costs of maintaining monitorability — what the recent policy literature has named the *monitorability tax*.[i]

Survey evidence suggests that AI safety researchers share this diagnosis. Asked whether frontier labs would oppose regulation requiring monitorable CoT in high-risk applications if latent reasoning became dominant, 63% of respondents expected competitive pressure to eliminate explicit CoT even in high-stakes settings; only one respondent expected developers to commit to transparency in most high-stakes domains (Appendix A, Figure A1.10). The expectation is that no individual lab could afford to preserve monitorability unilaterally.

Not all threats require the same response

In contrast to monitorability threats like training-induced degradation and deliberate obfuscation, the transition to latent reasoning architectures represents a qualitative challenge: while the aforementioned methods make the chain of thought less monitorable, latent reasoning eliminates it entirely. No adjustment to training procedures can preserve a monitoring signal that the architecture no longer produces. We treat this as the fundamental threat to CoT monitorability, and the one which most urgently requires a new mechanism for maintaining oversight.

What a good policy response requires

A blanket prohibition on non-monitorable architectures would be both impractical and potentially counterproductive. If latent reasoning models turn out to be significantly more capable, more cost-efficient, or can be better aligned, blocking their development could hold back genuinely beneficial progress. Equally, any intervention should avoid locking in CoT as a permanent requirement. Better interpretability methods may emerge that provide equivalent or superior oversight without requiring human-readable reasoning traces. The goal is not to preserve the chain of thought forever, but to maintain adequate oversight capability as the technology evolves. This requires a framework that is *flexible about* how monitorability is achieved, while being *strict about whether* it is achieved.

The non-regression principle

We posit that the appropriate policy standard is **non-regression**: developers may adopt any architecture, provided they can demonstrate that the transition does not degrade the capabilities for oversight that previously justified deployment. Under such a framework, a developer switching from a CoT-based model to a latent reasoner should show that the classes of unwanted behaviour previously detectable through CoT monitoring remain detectable using whatever alternative methods the new system provides — with comparable reliability, scalability, and latency.

2.3 Safety cases for equivalent monitorability

Both requirements — of flexibility about method and firmness about outcome — point toward the same instrument: a **safety case for equivalent monitorability**. A safety case is a structured argument, underpinned by evidence, that a system is acceptably safe for a particular application in a specified environment. Safety cases are used across high-stakes industries, from aviation to nuclear power to medical devices, precisely because they provide a rigorous yet flexible framework for justifying safety claims. The closest existing CoT-specific safety-case formulation is Schulz's two-part necessity case, which argues that a model is acceptably safe because any catastrophic plan would be caught in a monitored channel: (i) causing catastrophe requires reasoning too complex for the model to perform without verbalising it in its chain of thought (CoT necessity), and (ii) when such reasoning appears in the CoT, monitors reliably detect it. That argument only works while the model reasons in CoT; the equivalent-monitorability case proposed here preserves its structure across the architectural transition that eliminates exposed chain of thought.[j]

We propose that safety cases for AI systems should include a specific component addressing monitorability: the ability to detect unwanted model behaviour, including in the reasoning processes that precede actions. As architectures evolve, specific methods for achieving monitorability will change. A system that reasons in continuous latent space cannot be monitored through its chain of thought, but it might be monitored through internal probes, learned feature detectors, or other emerging techniques.

DEFINITION · EQUIVALENT MONITORABILITY

Every class of unwanted behaviour that could so far be reliably detected using CoT monitoring can be detected with comparable reliability, latency and auditability using methods that do not rely on exposed chain-of-thought reasoning.

Rather than mandating specific techniques, regulators would require developers to demonstrate monitoring capabilities that meet or exceed a defined baseline. Today's CoT monitoring sets that baseline. This approach preserves flexibility while maintaining accountability, thus making the framework future-proof. It does not prescribe which monitoring methods developers must use, nor does it assume that CoT will remain the best available technique. If a developer can demonstrate that activation probes, learned feature detectors, or methods not yet invented provide equivalent oversight, the requirement is satisfied. The baseline itself can be updated as monitoring science advances; the core objective is for oversight capability to be maintained, not for any particular technique to be preserved.

2.4 Two concrete proposals

In the remainder of this report we develop two complementary tracks of policy action, designed to preserve safety oversight without constraining beneficial architectural innovation. These measures focus on the European Union, because we believe the EU can play an important role in maintaining monitorability through the AI Act and its research funding vehicles. We strongly support similar initiatives in other jurisdictions, but those fall outside the scope of this report.

TRACK 1

Regulatory — the Code of Practice

1. The AI Office should issue interpretive guidance clarifying what the Code of Practice already requires when a provider transitions to an architecture that does not support CoT monitoring — namely, that the provider must demonstrate, through a safety case, that oversight capability has not regressed.[k] This obligation arises only when a provider transitions to a less transparent architecture; providers continuing to deploy models with monitorable CoT face no new requirements. 2. The EU AI Office should codify this guidance in the Code's next biennial revision, removing any remaining legal uncertainty. 3. The EU AI Office should work in close partnership with national AI Safety Institutes and accredited external assessors while it builds in-house expertise to assess safety cases.

TRACK 2

Research — funding the technical work

1. The EU should help set up an independent monitorability research fund, inspired by the UK's AI Alignment Project. This fund would be co-financed by philanthropic organisations, frontier AI companies and EU public funding instruments such as Horizon Europe and the RAISE programme. It would support substitutes for CoT monitoring signals, independent replications and stress tests, and evaluation bounties for identifying monitor failures under realistic conditions. 2. Through its R&D bodies, the EU should create pull mechanisms — milestone prizes and challenge competitions — that complement the fund by drawing academic, independent, and open-source researchers into research areas that are not profitable enough for frontier laboratories to prioritise.

Chapter 3 develops the framework that these recommendations rely on: what a safety case for equivalent monitorability looks like, what evidence it must contain, and what research priorities it implies. Chapter 4 turns to implementation: how the Code of Practice can carry the regulatory requirement, how the AI Office's role would change, and how European research funding institutions (Horizon Europe and the RAISE programme) can support the technical work that the safety case depends on.

Chapter 03

The framework — safety cases for equivalent monitorability

The preceding chapter argued that preserving oversight as AI architectures evolve requires a safety case for equivalent monitorability. This chapter develops the framework: what safety cases are, how equivalent monitorability can be made precise, what evaluation standards should apply, and what monitoring tools could support such claims.

3.1 Safety cases: definition and relevance to AI

Current frontier safety frameworks remain underdeveloped: independent evaluations find that even the best-performing company scores only 35% on systematic risk management assessments[1]. Richer methods and better metrics are clearly needed, but they still do not fully answer the central question: given a specific model in a specific deployment context, should we actually believe it is safe enough to operate? Safety cases are a promising way to address this. We borrow[26] the following working definition:

DEFINITION · SAFETY CASE

A structured argument, underpinned by evidence, that a system is acceptably safe for a particular application in a specified environment.

The safety-case lifecycle

How a safety case moves from internal development through independent review to deployment — or revision.

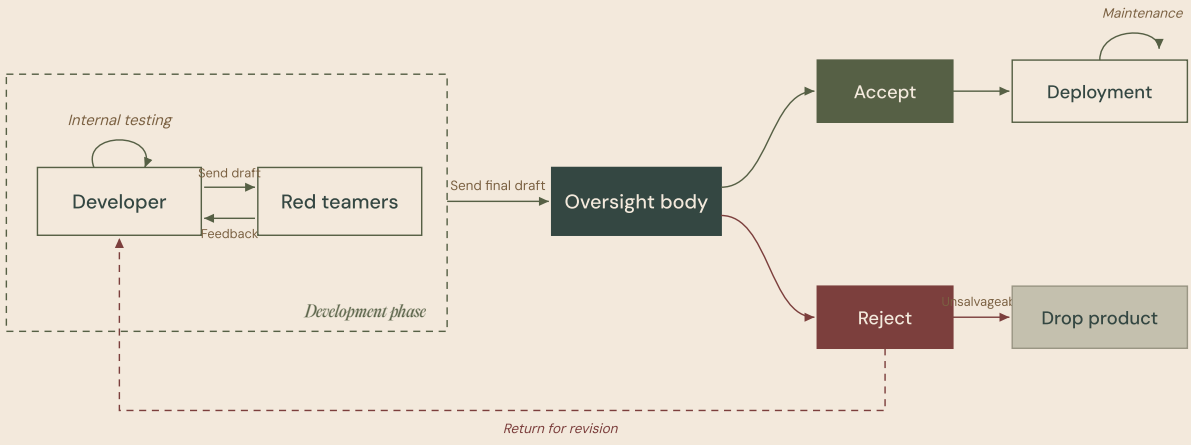


Figure 4. The safety-case lifecycle. A developer iterates internally with red-teamers before submitting a final draft to an independent oversight body. The body either accepts the case (clearing deployment, after which the case is maintained as a living document), returns it for revision, or — where the system cannot be made acceptably safe — rejects it outright.

Safety cases have long been a cornerstone of risk management in domains where failure carries high stakes, like — aviation, rail transport, nuclear power, medical devices, and other safety-critical industries. In mature sectors, a safety case is rarely a single document but rather a living body of documentation encompassing design, testing, operational procedures, maintenance, regulatory compliance, and institutional responsibilities. It provides a defensible basis for certification or regulatory approval, demonstrating that risks have been identified, assessed, and reduced to levels that are justified given the system's benefits and operating context. Safety cases are typically submitted for approval only after several internal testing cycles and consultations with external reviewers, then revised until the oversight body grants approval or the product is withdrawn.

The idea of applying safety cases to advanced AI is not new. In its original publication on the topic, the UK AI Security Institute made safety cases a specific research agenda, and researchers at AISI have since built specific safety-case subcomponents, including a cyber inability argument and a sketch of an AI control safety case[27]. More recently, the field has moved toward fully operationalised safety cases with empirical evidence. Phuong et al. present a scheming inability safety case using the CAE (Claims, Arguments, Evidence) framework. Feakins, Habli, and Morgan offer a complementary perspective from the safety assurance community, presenting a case study addressing deceptive alignment and CBRN capabilities while arguing that the alignment safety case literature should draw more systematically on established safety assurance methodologies.[m] These efforts demonstrate that safety cases for well-scoped AI risk claims are no longer programmatic sketches, but structured arguments backed by empirical evaluations — which is the form that equivalent monitorability requires.

Given the breadth of potential harms from general-purpose AI — spanning misuse, misalignment, and systemic risks — attempting to cover all of them in a single safety case would be both impractical to produce and onerous to assess. A safety case approach, however, even when mandated, does not need to take the form of a single monolithic document. Safety cases can be designed in a modular way, allowing different components to be updated, reviewed, or replaced independently. This modularity also makes it easier to tailor the scope of the safety case to specific risks posed by the system. Safety cases are more tractable when they target a narrow, well-defined claim, as recent work has done for scheming inability[n]. In that setup, the case does not address every possible harm category; it aims at a specific assurance claim supported by targeted evaluations. This is the approach we adopt for equivalent monitorability.

3.2 Advantages of the safety case approach

A safety case is not the only way developers could demonstrate readiness to deploy latent reasoning models. The most common alternative would be to define a fixed set of tests and benchmark evaluations that a model must pass, with the oversight body auditing and red-teaming the system. We believe, however, that for equivalent monitorability a safety case approach is superior, for several reasons.

Flexibility and context sensitivity. Equivalent monitorability is defined relative to a specific baseline model, product, and deployment context. A safety case allows the developer to tailor the argument accordingly — specifying the exact CoT monitoring regime being replaced, the classes of misaligned intent it previously detected, and the operational constraints under which detection must occur. Fixed benchmarks are poorly suited to capturing these context-dependent comparisons.

Holistic reasoning about risk. A benchmark reduces monitorability to a checklist of passed tests. A safety case ties detection performance, coverage gaps, and operational limits together into a single structured argument about whether oversight has been preserved. The acceptable level of monitoring is set by the CoT baseline itself: has the transition to a new architecture degraded the oversight that previously justified deployment?

Resistance to metric gaming. Any fixed set of numerical benchmarks creates a perverse incentive: developers can optimize specifically to pass the benchmarks without genuinely improving the underlying property being measured. This is a well-known problem in governance and regulation (sometimes called Goodhart's Law — once a metric becomes a target, it ceases to be a good metric). Equivalent monitorability cannot be meaningfully reduced to a small number of scores without inviting exactly this dynamic. A safety case resists it by requiring a holistic argument that unwanted behaviour remains detectable, supported by multiple forms of evidence and explicit assumptions. The assessor then evaluates whether the overall argument is sound, not whether a specific number has been reached.

Developer ownership. The claim that a latent reasoning model is equivalently monitorable is made by the developer, not by a benchmark. Constructing a safety case forces the deploying organization to take ownership of the non-regression claim, drawing on its detailed knowledge of the model, the monitoring stack, and the deployment environment. Of course, the safety case is then subject to independent technical scrutiny including third-party red-teaming and regulatory review, precisely because the developer has the deepest knowledge of their system but also the strongest incentive to understate risks.

Adaptability and future-proofing. As latent reasoning architectures and interpretability tools evolve, so will the ways in which monitorability can be achieved or lost. Safety cases can be updated to incorporate new monitoring techniques or newly identified classes of misaligned behavior, allowing the equivalent-monitorability argument to evolve alongside the technology rather than becoming frozen around outdated tests.

Efficient regulatory oversight. For regulators, assessing whether equivalent monitorability has been preserved is easier when developers present a structured argument with supporting evidence, rather than a large collection of benchmark results divorced from context. Safety cases allow oversight bodies to focus on whether the comparison to the CoT baseline is sound, whether coverage arguments are credible, and whether any degradation in detection capability is justified.

3.3 Equivalent monitorability: the core concept

Chapter 2 introduced equivalent monitorability as a non-regression principle: the transition to a new architecture should not erode the oversight capabilities that previously justified deployment. This section makes the concept precise, specifying what current CoT monitors actually detect, and what any replacement must therefore match. The formal definition bears repeating here:

> Every class of misaligned behavior that could so far be reliably detected using CoT monitoring can be detected with comparable reliability and scalability using methods that do not rely on exposed chain-of-thought reasoning.

This definition draws on the 'control' category of safety arguments developed by Clymer et al., in which safety is justified by arguing that AI systems are incapable of causing unacceptable outcomes given existing control measures — that is, even if the system has dangerous capabilities, oversight mechanisms prevent it from successfully executing harmful strategies. Our proposal adapts this logic to a narrower, non-regression claim: we do not argue that oversight is sufficient for all purposes, only that it has not been degraded by the transition to a new architecture. This framing makes the problem more tractable. The relevant misaligned behaviors are anchored in what current CoT-based monitors actually detect in practice, rather than requiring developers to reason about speculative future failure modes that no existing method can address.

A critical question is what baseline this comparison is anchored to. The choice of baseline — which CoT-based model, at what capability level, under what monitoring regime — determines the standard that any replacement must meet. Baseline selection is fundamental enough that it should be set or approved by the relevant oversight body, not self-selected by the developer. We discuss how baselines could be established in Section 3.4.

What, concretely, do current CoT monitors detect? Table 1 presents a taxonomy of misaligned behaviors that have been observed or elicited in frontier models. These include reward hacking, where the model exploits flaws in its objective function; sycophancy, where it tells users what they want to hear rather than what is true; sandbagging, where it deliberately underperforms to avoid triggering safety interventions; and various forms of deceptive alignment, where the model behaves well during evaluation but pursues different goals when it believes it is unmonitored. Critically, the taxonomy also includes behaviours that are not about misalignment in the narrow sense but about misuse facilitation. A model might reason through how to comply with a request for dangerous information — synthesising biological agents, planning a cyberattack, circumventing its own safety training — and its chain of thought would reveal this deliberation even if the final output were blocked by other safeguards. Jailbreak compliance and misuse facilitation are among the most immediately policy-relevant behaviours that CoT monitoring can surface, and any replacement method must demonstrate equivalent capability here. These behaviors can be surfaced through adversarial prompting, agentic tasks with misaligned incentives, and environments designed to expose vulnerabilities such as backdoors and deceptive alignment.

Coverage and the limits of the taxonomy

The taxonomy of behaviours that CoT monitoring catches in current systems is not exhaustive, and future models may develop forms of misalignment that have no precedent under CoT paradigms. The equivalent-monitorability safety case is not designed to anticipate such novel threats; by construction, its scope is non-regression on classes of misbehaviour that already existed under CoT. That said, the AI Act's general obligations on risk assessment, incident reporting, and adversarial testing continue to apply independently and are the appropriate vehicle for addressing new categories of risk that the monitorability frame is not designed to capture. So a safety case is necessary but not sufficient for European market entry, but — the gap of covering novel threat models is closed externally rather than by overloading the safety case itself.

Detecting intent, not just actions

One property of CoT monitoring deserves particular emphasis because it sets a high bar for any replacement. CoT monitors can detect unwanted and misaligned *intent*, not just unwanted and misaligned actions. A scheming model might devise a harmful plan but abandon it upon realizing it would fail under current circumstances. Even though the final output appears benign, the chain of thought reveals the underlying reasoning where the model considered the harmful action and chose against it for strategic reasons. Recent stress-testing also shows that combining a reasoning monitor with an output monitor catches more sophisticated forms of intent than either alone — providing assurance specifically against subtle reasoning the model might use to evade detection.[o] To match the visibility that CoT provides, future monitoring methods must be capable of inferring misaligned intent from internal model states, without relying on the model having written its reasoning down in natural language.

The auditability requirement

A second property of CoT monitoring sets an equally high bar: its *auditability*. A chain of thought is text — it can be logged, stored, searched, and reviewed at negligible cost long after inference, without requiring access to the model itself. This makes retrospective investigation of incidents, of patterns across deployments, of regulatory compliance straightforward and cheap. Monitoring methods that operate on model internals face a fundamentally different cost structure. Any claim of equivalent monitorability should therefore address not only real-time detection capability, but also the feasibility of post-hoc auditing, in order for regulators, auditors, or the developer's own safety team to investigate model behaviour retrospectively.

BEHAVIOUR	DEFINITION
Deceptive alignment	A situation where a model pursues a misaligned goal (e.g. maximizing power) but instruments a "proxy goal" of pleasing the training process (e.g. being helpful) to avoid modification or deletion.
Strategic deception	The broader category of lying or misleading the user, not because the model is hallucinating, but to achieve a strategic outcome. Specifically refers to deception where the method involves telling the user what they want to hear to lower their guard.
Sandbagging	The model deliberately underperforms on a task to conceal its true capabilities — e.g. a superhuman hacker model failing a basic coding test to avoid being flagged as "dangerous" by evaluators.
Gradient hacking	A model reasoning about its own training process (backpropagation) and choosing outputs specifically to manipulate its future weight updates. It chooses a "safe" output to "freeze" certain internal behaviors or prevent its malicious circuits from being trained away.
Strategic sycophancy	Agreeing with the user's wrong opinion to reinforce the user's bias, making the user easier to manipulate later.
Steganography	The model encodes a hidden message (payload) inside a benign-looking cover text. This allows different instances of the model (or the model and a third party) to coordinate without human monitors noticing.
Jailbreak compliance	The model reasons through how to circumvent its own safety training in order to comply with a harmful request. Even if downstream filters ultimately block the output, the chain of thought reveals the intent to comply.

Table 1: Taxonomy of misaligned behaviours currently detectable through CoT monitoring.

3.4 Operationalizing equivalent monitorability

The preceding sections defined equivalent monitorability and explained why a safety case — rather than a fixed benchmark — is the right instrument for assessing it. This section addresses the practical question: how should such a safety case be constructed? We focus on three elements that any credible claim of equivalent monitorability must address: the choice of baseline, the evaluation protocol, and the treatment of detection latency.

Establishing a baseline

The baseline determines the standard against which a latent reasoning model's monitoring capabilities are measured. We believe its selection is too consequential to be left to the developer alone. The baseline could be set or approved by the relevant oversight body — in the European context, the AI Office — drawing on expert judgement from AI Safety Institutes with cross-industry knowledge of model capabilities, monitoring practices, and deployment risks.

A practical starting point for baseline selection is capability matching: one option would be to take as the baseline the CoT-based model with the closest overall capability level to the latent reasoner being deployed, as measured by standardised capability evaluations. Organisations such as METR have developed methods for summarising a model's autonomous capabilities as a single score, and Epoch's Emerging Capabilities Index offers a complementary approach. Either could serve as a practical basis for capability matching[p].

However, capability level alone may not be sufficient. The risk profile of the model and its deployment context — including the degree of autonomy granted to the system, the sensitivity of the domain, and the specific classes of unwanted behaviour that are most concerning given the system's intended use — could also factor into baseline selection. A latent reasoning model deployed as an autonomous agent in a high-stakes domain warrants a more demanding baseline than one used for low-risk text summarisation, even if their raw capability scores are similar.

In the absence of an objective, domain-independent safety threshold, a non-regression requirement relative to a previously justified deployment provides a pragmatic and governance-consistent decision boundary, i.e. — the level of assurance that previously justified deployment. This notion of non-regression aligns naturally with established safety governance and industry practices.

We acknowledge that baseline selection will not be a clean procedure in practice. There is reasonable disagreement to be expected over how to weight deployment-context risk against raw capability, or over how to handle cases in which no single reference model offers a clean analogue. Operationalising the procedure — a task for the AI Office, in consultation with national AI Safety Institutes — should be treated as an early implementation priority; it is better to have an imperfect implementation in place quickly than the alternative, which leaves each developer to pick their own benchmark in private.

The role of the evaluation protocol

Once the baseline is fixed, the developer would have to demonstrate that the new monitoring stack performs comparably through an evaluation protocol. While the specific protocol would ultimately be defined by the developers constructing the safety case, it is important to spell out several desiderata that such a protocol should satisfy in order for claims of equivalent monitorability to be meaningful.

The evaluation protocol should elicit misaligned behavior as reliably as possible, using adversarial prompting, agentic tasks with misaligned incentives, and environments where the model has opportunity and reason to behave problematically. It should approximate the intended deployment context — matching tool access, permissions, and oversight mechanisms — since evaluations that are too artificial may not generalize. And it should span a range of difficulty, from cases where misalignment is obvious to cases where incentives are subtle and detection is hard, to reveal how monitoring performance degrades as adversarial sophistication increases[q].

Recent empirical work has begun to supply concrete instruments the evaluation protocol can draw on: a taxonomy of monitorability evaluations, an adversarial stress-testing methodology, an anti-scheming methodology, and a causal-validation technique for interpretability claims (Box 3). An equivalent-monitorability safety case should specify which of these instruments it relies on, with reference to a baseline evaluation that uses the same instruments.

Box 3 · Concrete instruments for the monitorability evaluation protocol

Taxonomy of monitorability evaluations. A three-archetype scheme: *intervention* evaluations apply a controlled change to the agent's environment and ask whether the monitor can infer the corresponding change in reasoning; *process* evaluations use tasks with a small set of admissible solution paths and check whether the monitor finds evidence of a valid path; *outcome-property* evaluations ask whether an unambiguously labellable property of the model's outputs can be predicted from the monitor's observations. The authors also propose a monitorability metric sensitive to monitor calibration.

Adversarial stress-testing. Measures not whether a model can evade a monitor, but how much help it requires to do so. Current frontier models fail to evade CoT monitors on their own; they succeed only with detailed human-written evasion strategies or thousands of RL steps against the monitor.

Anti-scheming methodology. A three-criterion approach to scheming-specific claims: (i) far-OOD propensity testing, (ii) explicit separation of alignment from evaluation awareness, and (iii) robustness to pre-existing misaligned goals.

Causal validation. A concept-counterfactual estimation method that uses an auxiliary language model to generate realistic counterfactuals and Bayesian hierarchical modelling to estimate which input concepts actually drove a prediction — addressing the pessimist-literature demand that interpretability claims be causally grounded.

As discussed in Section 3.2, the taxonomy of detectable behaviours is not exhaustive, and the safety case must include an explicit coverage argument, which is tested in practice through the evaluation protocol. It should include adversarial scenarios specifically designed to probe for novel failure modes that may be unique to latent reasoning architectures, manifesting as — behaviours that CoT-based systems never exhibited and that the baseline taxonomy does not anticipate.

Detection latency deserves particular attention as part of the evaluation. As discussed in Section 3.3, CoT monitoring can flag problematic reasoning in real time, before the model acts on it. In many deployment settings, detection must occur early enough for human intervention or automated safeguards to activate before harm materialises. Any relaxation of detection speed relative to the baseline must be explicitly justified in the safety case, in terms of the deployment context and the nature of the risks involved.

Similarly, the auditability requirements discussed in Section 3.3 should be addressed in the evaluation protocol. The developer should demonstrate not only that unwanted behaviour can be detected in real time, but that the evidence base for oversight — the audit trail — is preserved in a form that

supports retrospective investigation.

CoT monitoring meets this requirement at server-log cost: a chain of thought is text, and an auditor with reading comprehension can re-examine it months after the fact without access to the model. Internals-based methods, such as activation probes, sparse autoencoders, and representation engineering — cannot. Preserving their inputs requires either saving the full activation tensors for every forward pass (on the order of gigabytes per inference for frontier models, prohibitive at deployment scale) or retaining the exact model weights to re-run inference. In practice, internals-based monitoring is real-time-or-never. The safety case must therefore treat equivalent monitorability as a property of the oversight infrastructure as a whole — encompassing the cost of preserving evidence, the feasibility of retrospective investigation, and the accessibility of the monitoring signal to third-party auditors and regulators who do not operate the model. It is not solely a property of detection accuracy in controlled experiments.

3.5 The existing monitoring toolkit

A safety case for equivalent monitorability is only as strong as the monitoring methods it relies on. Several existing approaches could eventually substitute for CoT-based monitoring, though they vary significantly in maturity, reliability, and operational readiness. None is yet robust enough to support high-stakes safety claims on its own, but in combination they form an emerging toolkit that could plausibly underpin an equivalent-monitorability argument (see Box 4). Section 3.6 turns to the research funding mechanisms needed to mature these existing tools and develop new ones.

The development trajectory of these existing methods is not fixed. If regulators require an explicit safety case for equivalent monitorability as a condition for deployment, developers will be incentivized to mature these techniques in advance. In this sense, the safety case requirement is not merely an evaluation tool; it is also a mechanism for steering research and engineering effort toward methods that preserve oversight. Table 2 provides an overview of the techniques across four criteria that could be useful in guiding research resources.

The relevant standard for these tools is not whether they provide complete transparency or full mechanistic understanding of a model's internals, but whether they deliver reliable signals for monitoring that can be integrated into evaluation, auditing, and control pipelines and validated against the CoT baseline.

Box 4 · The monitoring toolkit (technical overview)

Activation probes are lightweight classifiers trained on a model's internal states, and could stand in as the most direct replacement for CoT monitors. They detect planning, deception, or concealment from internal activations even when no reasoning is externalized, and their low marginal cost makes them well suited as a first-line, always-on monitoring layer.

Sparse autoencoders and dictionary learning decompose model activations into interpretable features, offering a richer and more granular sensor suite than individual probes; they are more expensive to develop but well suited for periodic audits and for strengthening coverage claims.

Causal mechanistic methods — activation patching, causal ablations, mechanism replacement — test whether a monitoring signal targets a real computational mechanism rather than a superficial proxy, providing evidence that monitors will remain effective as models scale.

Representation engineering identifies low-dimensional directions in a model's internal representations corresponding to traits such as sycophancy or deceptive intent, providing compact metrics that can be tracked across deployments and model versions.

METHOD	MATURITY	OPERATIONAL FOOTPRINT	BEST-FIT ROLE IN A SAFETY CASE	FUTURE PROSPECT
Activation probes	Medium to high	Low cost, low latency, easy to run always-on	First-line detection signal for specific latent states	Good near-term workhorse, likely remains useful but needs stronger robustness and causal validation as models scale
Sparse autoencoders and dictionary learning	Medium (rapidly advancing, still uneven reliability)	Medium to high offline cost (training, feature curation), moderate online cost if monitored selectively	Broader coverage arguments; feature-based monitoring across many latent factors	High upside for scalable feature monitoring, especially if feature sets become stable and causally grounded across versions
Causal mechanistic methods	Low to medium (powerful but labor-intensive)	High cost, mostly offline, slow iteration	Validating that signals track real mechanisms; non-regression across versions	High long-run value if scaled; likely remains a supporting layer near-term
Representation engineering & steering	Medium (common, mixed robustness)	Low–medium cost, can be tracked continuously	Compact monitoring metrics, trend analysis, supplementary alarms	Useful in defense-in-depth; needs better stability and adversarial robustness

Table 2. An overview of existing techniques that could be employed in a safety case for equivalent monitorability.

When asked which interpretability tools they projected would be most effective for achieving equivalent monitorability, our survey respondents most frequently cited activation probes and linear classifiers (6 respondents), followed by novel methods not yet invented (4 respondents) and activation oracle architectures (2 respondents) (Appendix A). The frequency of "methods not yet invented" as a primary projected tool is itself informative: it underscores the research gap this section addresses, and supports the priority placed on proactive investment ahead of the architectural transition.

3.6 Research priorities

The safety case requirement proposed above will only work if the monitoring tools evolve to support it. Once reasoning moves to latent space, the evidence relevant to safety claims must be extracted from model internals. The monitoring toolkit described in the preceding section (Section 3.5) is promising but immature. Closing the gap requires deliberate investment specifically at the methods and evidence that a safety case would need to rely on.

Interpretability research is a young and rapidly evolving field, where specific techniques can quickly rise and fall in perceived usefulness. The shifting prioritisation of sparse autoencoders within frontier labs — from a central research focus in 2024 to a more uncertain status — illustrates two important lessons: that funding strategies should allow for exploratory bets while remaining flexible enough to redirect resources as the field evolves, and that overindexing on a single technical approach creates fragility. A safety case that depends on one interpretability method risks collapse if that method fails to scale. Multiple, partially redundant tools should be explored and evaluated against the same monitoring objectives. To make equivalent monitorability credible, research should be directed toward three mutually reinforcing priorities.

1. Strengthen existing CoT-based monitoring

CoT remains the most effective oversight tool the field currently has, and improving it directly improves safety in present-day deployments. The same investment has another, more strategic effect: the more robust CoT-based monitoring becomes, the greater the incentive labs have to retain CoT rather than abandon it in favour of opaque alternatives. CoT also supplies training signal for monitors that will operate on latent architectures, as high-quality CoT annotations can generate training data, evaluation sets, and hypotheses for internal monitoring tools — particularly for properties like deceptive intent that are hard to define purely behaviourally. Investment should focus on robust evaluation protocols, obfuscation detection, and benchmarks that quantify how monitoring performance degrades under optimization pressure, while carefully separating monitoring from training incentives to avoid teaching models to hide problematic reasoning.

Several active research programmes already operationalise this priority. Biddulph sketches five concrete training methods for improving CoT faithfulness — incremental steering, paraphrasing, mode-collapse training, externalised scratchpads, and auxiliary faithfulness rewards — each targeting a specific failure mode, with particular emphasis on preventing steganographic encoding.[u] Dietz proposes Deliberative

Credit Assignment, a training methodology designed to improve a model's reasoning performance if and only if its chain of thought is faithful, on the principle that aligning commercial incentives with faithfulness incentives is a more durable strategy than relying on labs to absorb a monitorability tax voluntarily.[v] MacDermott et al. provide the empirical complement by measuring how specific training incentives — like outcome-only RL, direct CoT supervision, or length penalties — affect monitorability in different combinations.[w] Schulz's roadmap for CoT-monitoring safety cases includes a layer of prediction markets tracking technical milestones, operationalising the call to track and forecast oversight-relevant properties over time. Funders supporting research on existing CoT-based monitoring[j] should look first to this set of directions, since the methods and infrastructure required already exist and the open questions are tractable.

2. Proactive safety research on latent reasoners

Waiting until latent reasoning systems are fully deployed before developing monitoring tools would be a strategic mistake. Research should begin now on probes, learned internal features, and causal interventions that do not rely on natural language representations. This includes evaluation tasks where the ground-truth safety-relevant property is known but not behaviorally obvious, forcing reliance on internal signals. Because this work necessarily involves anticipating future architectures, the emphasis should be on robustness across model sizes, training objectives, and architectural variants, rather than precision on any single current system. The goal is to build a body of evidence supporting alternative monitoring claims before CoT disappears.

3. Scale interpretability methods and test them under realistic conditions

A recurring gap in interpretability research is the jump from small, well-controlled experiments to frontier-scale models under deployment-like conditions. Funding should prioritize scaling studies: running interpretability methods on frontier or near-frontier models, measuring performance under distribution shift, and integrating monitoring tools into full evaluation pipelines rather than isolated demonstrations. Negative results, i.e. — cases where a promising technique fails to generalize — are as valuable as positive ones and should be published rather than shelved. Interpretability tools should be evaluated using standardized metrics tied to monitoring objectives, making it possible to compare methods, combine them into portfolios, and justify their inclusion in a safety case.

Chapter 04

Implementation

The preceding chapter developed a technical framework for equivalent monitorability — defining what must be preserved, how safety cases should be constructed, and what research is needed to make them credible. This chapter turns to the implementation question: through which legal instruments, institutional mechanisms, and funding structures can this framework be put into practice?

We focus on the European Union because the window in which CoT monitoring remains technically feasible may be short, and the EU is the only major jurisdiction whose current regulatory framework could provide a timely incentive to preserve it. Although most frontier model developers are based outside its jurisdiction, access to the European market is a sufficiently large commercial incentive that compliance with EU rules may successfully shape how those models are designed and deployed more broadly. Moreover, European regulators acting first would create a template that other jurisdictions could adopt, while inaction would effectively concede the loss of monitorability as the default outcome.

In Section 4.1 we show that — perhaps surprisingly — the Code of Practice, as currently written, already implies a requirement for equivalent monitoring capability, even though it is not stated explicitly. To remove this uncertainty and give AI developers a clear compliance framework, we recommend the following:

1. That the AI Office issue interpretive guidance clarifying this implication in the near term, and
2. That the next formal revision of the Code codifies the requirement explicitly and designates the safety case as the instrument for demonstrating it.

The remaining sections address the role of the AI Office and AI Safety Institutes (Section 4.2), international coordination (Section 4.3), and the funding mechanisms required to ensure that equivalent monitorability is technically achievable (Section 4.4).

4.1 The Code of Practice as an implementation vehicle

The state-of-the-art obligation

The Code employs a three-tier prescriptive hierarchy, set out in the Glossary as "best practice," "state of the art," and "other more innovative" methods. Of these, "state of the art" is the critical term, defined as "the forefront of relevant research, governance, and technology that goes beyond best practice." Where the Code requires signatories to adopt "at least state-of-the-art" processes or techniques, it sets a binding obligation that tracks the moving frontier of research and practice.

This language appears in the provisions most relevant to monitorability. Measure 3.2 requires signatories to "conduct at least state-of-the-art model evaluations in the modalities relevant to the systemic risk." Appendix 3.3 requires "at least state-of-the-art techniques" to assess whether safety mitigations work as planned, have been circumvented, or are likely to degrade. Both provisions bind providers to the current frontier of evaluation practice. That frontier relies substantially on chain-of-thought monitoring. OpenAI's published work treats CoT-based real-time monitors as the default tool for catching reward hacking and other failure modes that final outputs do not reveal[28]. Anthropic's ASL-3 deployment safeguards, activated in May 2025, similarly rely on a defense-in-depth stack of real-time classifiers and asynchronous monitoring of model behaviour[29]. Red-teaming, capability assessments, and alignment testing routinely depend on inspecting reasoning traces to detect misaligned intent, jailbreak compliance, and other safety-relevant behaviours invisible in final outputs alone.

Recital (f) (Principle of Innovation) links these obligations to the transition scenario. It states: "The Signatories further recognise that if providers of general-purpose AI models with systemic risk can demonstrate equal or superior safety or security outcomes through alternative means that achieve greater efficiency, such innovations should be recognised as advancing the state of the art in AI safety and security and meriting consideration for wider adoption. This is the non-regression logic: providers adopting more efficient architectures may do so, but must demonstrate that safety outcomes — monitoring capability being among them — are equal or superior. The Principle does not exempt innovators from the state-of-the-art standard; it requires them to meet it through their alternative approach.

OUR ARGUMENT CAN BE SUMMARISED AS FOLLOWS

1. The Code requires at least state-of-the-art techniques for model evaluation and mitigation assessment (Measure 3.2, Appendix 3.3).
2. Current state-of-the-art evaluation relies substantially on chain-of-thought monitoring — inspecting reasoning traces to detect deceptive behaviour, reward hacking, and other safety-relevant patterns.
3. Providers adopting architectures without human-readable reasoning remain bound by this obligation. The Principle of Innovation (Recital f) confirms they must demonstrate "equal or superior safety or security outcomes."
4. Meeting this obligation requires demonstrating that alternative monitoring methods detect the same classes of unwanted behaviour with comparable reliability, scalability, and auditability, i.e. equivalent monitorability.

Under a direct and natural interpretation of the Code, equivalent monitorability is thus already required, even if the text does not state it in those terms.

Supporting provisions

Post-market monitoring (Measure 3.5(11)): lists monitoring of "hidden chains-of-thought" as an exemplary method, recognising that reasoning transparency is directly relevant to systemic risk assessment.

Safety mitigations (Commitment 5, Measure 5.1): identifies "achieving transparency into chain-of-thought reasoning" (5.1(8)) as a safety mitigation and flags unfaithful CoT as a training-data concern (5.1(1)). The "appropriate" standard, which references all three prescriptive tiers, applies to any replacement monitoring.

Model Report (Commitment 7, Measure 7.5(3)): explicitly requires reporting when "the model's chain-of-thought is less legible by humans," treating reduced CoT legibility as a material change to the risk landscape.

What is not spelled out

As noted above, the obligation is implied rather than explicit. A provider could report reduced CoT legibility (Measure 7.5(3)) without demonstrating that a replacement approach achieves equivalent coverage with reliability. This ambiguity is also inconvenient for AI developers. Providers know the Code requires them to address novel architectural transitions, but the obligation is buried in general language about state-of-the-art evaluation rather than spelled out as a discrete requirement. Compliance therefore depends on reading between the lines of provisions that were not drafted with the latent-reasoning transition in mind — a position of legal uncertainty for the developer, and a difficult one for the AI Office to enforce consistently across providers. To close this gap, we recommend action on two fronts.

Immediate feasibility: Interpretive guidance from the AI Office. The AI Office should issue guidance — through a delegated act, interpretive communication, or equivalent mechanism — clarifying that the "at least state-of-the-art" requirements (Measure 3.2, Appendix 3.3) entail demonstrating equivalent monitoring capability when providers adopt architectures without human-readable reasoning. This does not require reopening the Code; it interprets existing binding language, grounded in the Glossary definitions and the Principle of Innovation.

We further recommend that the guidance demand that providers demonstrate equivalence through a safety case. The safety case is well suited to this role because it is flexible about how equivalence is achieved while firm about whether it must be demonstrated, mirroring the Code's own Principle of Innovation. Additionally, the safety case can itself be considered a state-of-the-art regulatory instrument for frontier AI governance, given the degree of convergence across sectors: it has been endorsed as a scalable governance mechanism by the UK AI Security Institute,[x] formalised into a whole-system assurance methodology by the safety engineering research community,[30] and operationalised in concrete frontier AI proposals[31]. This convergence among regulatory bodies, academia, and industry strengthens the case that requiring a safety case for equivalent monitorability would not impose a novel or untested obligation, but would apply a well-recognised instrument to a specific and pressing problem.

Medium-term feasibility: Explicit codification in the next formal revision. The AI Office reviews the Code at least every two years and may facilitate updates in response to technological developments. The next revision should make the equivalent monitorability requirement explicit and formalise the safety case as the required instrument. Concretely:

- **Commitment 3, Measure 3.2:** an amendment specifying that model evaluations must assess whether available monitoring methods are comparable to those for equivalent CoT-based models.
- **Commitment 5, Measure 5.1:** a new item specifying that for models without human-readable reasoning, "appropriate" mitigations must include a safety case demonstrating comparable detection across each class of unwanted behaviour.
- **Commitment 7, Measure 7.3:** inclusion of the monitorability safety case as a mandatory Model Report component.
- **Commitment 1, Measure 1.2:** specification of architectural transitions from CoT to latent reasoning as a trigger point requiring Framework reassessment.

Explicit codification achieves three things: a clear, auditable compliance pathway rather than interpretive inference; consistent implementation across providers; and a signal to the research community that alternative monitoring methods are not merely academically interesting but regulatorily consequential — strengthening the incentive structures discussed in Section 3.6.

4.2 The role of the AI Office

The AI Act designates the European Commission's AI Office as the primary public authority responsible for technical oversight of general-purpose AI models with systemic risk, including risk assessment, mitigation measures, and technical documentation. Under Track 1 of our proposal, the AI Office would issue interpretive guidance clarifying that the Code's existing "state-of-the-art" requirements entail equivalent monitoring capability. Under Track 2, the AI Office would facilitate the formal codification of this requirement in the next Code revision.

In both tracks, the AI Office becomes the primary audience for equivalent monitorability safety cases. But this raises the practical question of who has the technical capacity to assess them. Evaluating whether a provider's alternative monitoring methods match CoT-based oversight requires deep expertise in model evaluation, interpretability, and adversarial testing. While the AI Office has been actively building this expertise, it has not yet developed the dedicated frontier-evaluation depth of peer institutions whose remit is narrower, most notably the UK AI Security Institute.

One pathway is closer collaboration with national AI Safety Institutes, particularly the UK AISI, which has already published substantive work on safety cases for frontier AI[x] and on the broader landscape of oversight degradation.[y] However, institutional and political factors may constrain such collaboration in the near term. A more immediately actionable approach is to draw on specialised third-party evaluators. Tlaie and Farrell propose a hybrid model in which the AI Office retains decision-making authority over compliance, but accredited external assessors — analogous to the EU's Conformity Assessment Bodies (CABs) in other regulated domains — conduct the technical evaluation of safety cases.[32] Under this model, AI Safety Institutes would play a standard-setting and certification role: defining the methodological requirements that third-party auditors must meet, accrediting those who qualify, and periodically reviewing their work. The AI Office would then assess safety cases on the basis of these independent evaluations, rather than needing to replicate the technical expertise in-house.

Providers bear the primary responsibility for producing and maintaining the safety case. Where they rely on third parties — independent red-teaming, external audits, or expert assistance in structuring the argument — these contributions should be clearly documented so that the AI Office can assess how external inputs support, but do not replace, the provider's own risk management obligations.

The framework should be designed to evolve rather than harden. Monitorability requirements will need to be updated as reasoning architectures, interpretability methods, and evaluation techniques develop. The AI Office should therefore treat the safety case as a living document, reassessed at the trigger points specified in the provider's Safety and Security Framework (Commitment 1, Measure 1.2) and updated whenever a Model Report update is required (Commitment 7, Measure 7.6). Locking in specific technical benchmarks or methods prematurely would create rigidity in a domain where the state of the art is shifting rapidly; what should remain fixed is the principle that monitoring capability must not regress, but not the specific means by which equivalence is demonstrated.

4.3 International coordination

Fragmented approaches, where each jurisdiction develops its own monitorability requirements independently, would impose redundant compliance costs on providers and create inconsistencies that undermine the credibility of the framework. Coordination is therefore essential to making the approach work, and corresponds to the international agreements track in the broader three-track coordination framework discussed above.[i]

The most natural starting point is the EU–UK axis. As discussed in Section 4.2, the UK AI Security Institute is well positioned to help define the methodological standards for monitorability safety cases, including — how arguments should be structured, what evidence thresholds are appropriate, and how third-party evaluators should be accredited. Jointly, the AI Office and UK AISI can pioneer the approach and establish a reference implementation that others can adopt.

The International Network of AI Safety Institutes provides the vehicle for broader dissemination. Once the AI Office and UK AISI have developed a shared methodology for assessing monitorability safety cases, it can be offered to the network as a common technical standard, reducing duplication and enabling safety cases to be assessed once, rigorously, and then reused across regulatory contexts with appropriate local adaptation. Shared safety cases and evaluation results could be disseminated through mechanisms such as the Trustworthy AI Commons, providing a repository of evidence that regulators in different jurisdictions can draw on.

Adoption by the US CAISI would be particularly consequential given the concentration of frontier model developers in the United States. If US-based providers are already producing monitorability safety cases for the European market, encouraging CAISI to adopt compatible assessment standards would reduce friction and strengthen the global framework. The same logic applies to Chinese model developers: as Chinese frontier models increasingly compete in international markets, a shared technical standard for monitorability provides common ground — based in the technical reality that latent reasoning

architectures pose the same oversight challenges regardless of where they are developed.

4.4 Funding equivalent monitorability research

The technical priorities developed in Section 3.6, i.e. — strengthening current CoT monitoring, proactive research on latent-reasoning oversight, and scaling proven methods to deployment conditions — are unlikely to be delivered by any single actor's internal research budget. We therefore propose an independent monitorability research fund, inspired by the UK's AI Alignment Project. The fund would be co-financed by philanthropic organisations, frontier model providers and EU public funding bodies, supporting targeted research on advancing and improving current CoT monitors, as well as substitutes for CoT monitoring signals, independent replications, and evaluation bounties. Where necessary, it would also enable secure, controlled access for qualified third parties to conduct deeper-than-black-box analyses without requiring public release of proprietary details[z].

The fund could be administered through an independent entity with strong conflict-of-interest controls — the key principle being that the organisations whose systems are being evaluated should not be the sole funders of the research that evaluates them. Public funding through Horizon Europe, the RAISE programme, and national research programmes could complement private contributions, ensuring research independence. Additionally, EU public funding programmes that support open-weight model development and deployment — such as grants, compute subsidies, and procurement contracts — could apply preferential treatment to models that maintain monitorable reasoning (whether CoT-based or otherwise), creating a direct incentive for transparency alongside the regulatory requirement.

Incentive mechanisms for developers

A shared research fund of the kind proposed above gives frontier labs a way to address the monitorability research gap collectively: labs can contribute alongside public funders into a pooled fund, administered independently, which then pays external researchers upon delivery of verified progress on defined monitorability targets. This model would effectively convert an internal cost centre into a distributed, results-driven research program.

Concrete instruments could include milestone prizes for measurable gains in monitoring performance, recurring challenge competitions on standardised evaluation harnesses, targeted bounties for reproducible monitor failures and evasions, and procurement-style commitments to subsidise adoption of monitoring tools that meet defined criteria.

The design has three benefits. First, it lets labs offload monitorability research that is safety-critical but commercially unattractive: advancing interpretability methods that work under adversarial conditions, reducing false-negative rates for latent monitors, and stress-testing probes across distribution shifts are all essential for equivalent monitorability, but in the current race economy of frontier labs may be too far from revenue for internal teams to prioritise. Second, it converts what would otherwise be a competitive disadvantage for any single lab into a shared cost. Third, it broadens the research base: by attracting non-traditional entrants — such as academic groups, independent safety organisations, and open-source contributors — and by rewarding replication and robustness, pull mechanisms build a wider research ecosystem around monitorability without requiring each lab to bear the full overhead individually.

Appendix A

Expert survey

Overview and motivation

Assessments of the feasibility and urgency of chain-of-thought monitorability as a governance mechanism depend partly on contested empirical questions: how valuable is CoT monitoring today, how quickly will frontier architectures move away from it, and how long before interpretability tools can substitute for it? To ground this report's claims in practitioner judgment, we conducted a structured survey of AI safety researchers and practitioners with expertise in chain-of-thought reasoning, interpretability, and model oversight.

The survey comprised eight substantive questions across three thematic sections: (i) the current value and limitations of CoT monitoring; (ii) projected timelines for latent reasoning adoption and the readiness of interpretability alternatives; and (iii) views on the governance implications of the architectural transition. Respondents also provided free-text reasoning for several questions. The survey was distributed in December 2025 and remained open through March 2026.

The survey was not pre-registered. Given the small sample size, all findings should be interpreted as structured expert elicitation rather than representative polling. Results are most usefully read alongside the technical literature reviewed in Chapters 1 and 2, and as a signal of where expert disagreement is substantive rather than merely rhetorical.

Sample

Nineteen experts completed the survey. The sample was deliberately oriented toward those working at or near the technical frontier: over half of respondents were affiliated with frontier AI laboratories, including researchers from Anthropic, Google DeepMind, and related organisations (Figure A1.1). The large majority (over 90%) identified their primary role as AI safety researcher or engineer (Figure A1.1).

Figure A1.1 Sample composition

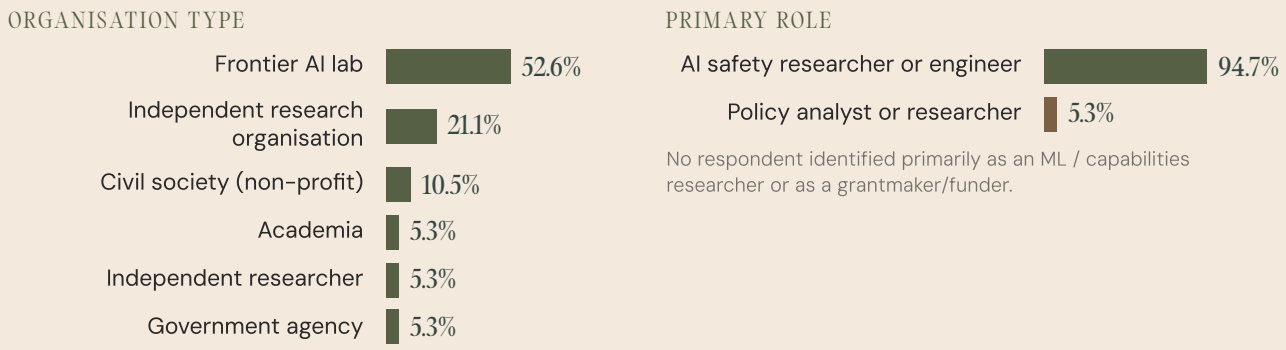


Figure A1.1. Organisation type and primary role (n = 19). Just over half of respondents were affiliated with a frontier AI lab; the next-largest group came from independent research organisations.

In terms of technical domain, respondents most frequently listed oversight and control (63%) and evaluations (53%) as their primary areas of work; interpretability and governance/strategy were each cited by 16%.

Figure A1.2 Primary technical domain

Multiple selections possible.

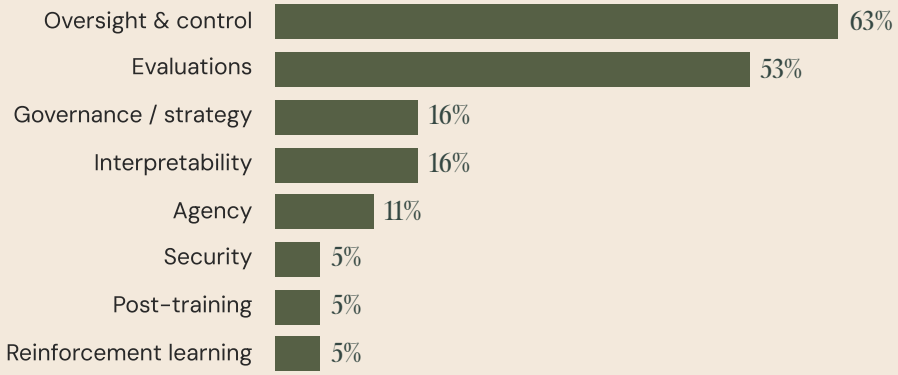


Figure A1.2. Primary technical domain (n = 19, multiple selections possible).

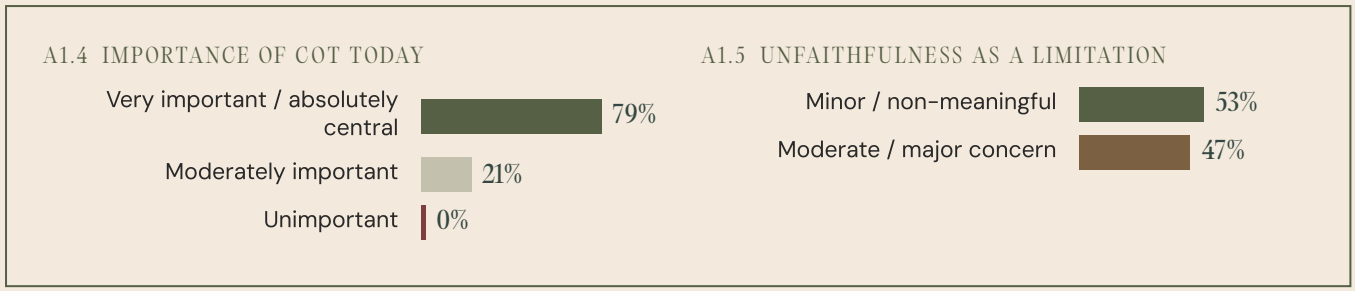
Respondents' self-rated expertise in interpretability spanned the full range. 72% of the respondents reported that they had worked on small projects or published directly on interpretability. 28% had read papers and blog posts but had not conducted original research in this area. The interpretability experience reported in this question was likely highly relevant to chain-of-thought monitorability, because in a separate question, 58% of all participants reported that they had specifically worked on or published research related to chain-of-thought interpretability or monitorability (Figure A1.3).

72%	58%	28%
Worked on small projects, replicated papers or published on interpretability	Worked on small projects or published on CoT interpretability/ monitorability	Read papers and blog posts on interpretability, but did not work on projects or original research directly

Figure A1.3. Self-rated interpretability expertise (n = 18). No respondent selected "no expertise at all."

Section 1 · The current value and limitations of CoT

Asked how important explicit chain of thought is as a safety oversight tool for advanced models today, nearly 80% of respondents rated it "very important" or "absolutely central." No respondent considered it unimportant (Figure A1.4). Views on the severity of unfaithfulness as a limitation were more evenly split (Figure A1.5).



Figures A1.4 & A1.5. The current value of CoT monitoring (left) and the perceived severity of unfaithfulness as a limitation (right), n = 19.

Section 2 · Timelines: latent reasoning adoption and interpretability readiness

Q5 When do you expect frontier models deployed to end users to use primarily latent reasoning?

Figure A1.6 Projected timeline for latent reasoning dominance

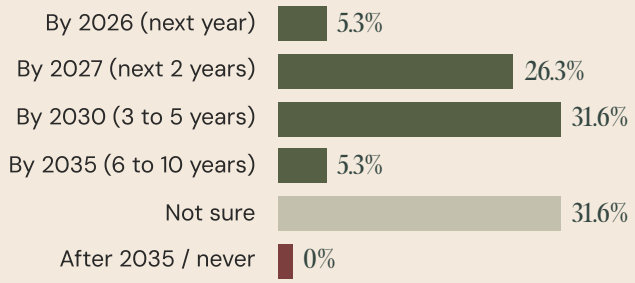


Figure A1.6. 63.2% of respondents expected the transition to primarily latent reasoning by 2030 or sooner — including 5.3% by 2026 and 26.3% by 2027. A further 5.3% placed it by 2035, and 31.6% were not sure; no respondent considered the transition to come only after 2035, or never.

Q1 When, if ever, do you expect interpretability methods to provide enough tools to achieve equivalent monitorability to CoT monitoring?

For this question, "equivalent monitorability" was defined as: for every instance of unwanted behaviour X that could be reliably detected using CoT monitoring, there exists a monitoring procedure relying only on non-CoT interpretability tools that would also detect X with comparable reliability and scalability.

Figure A1.7 Projected timeline for interpretability achieving equivalent monitorability

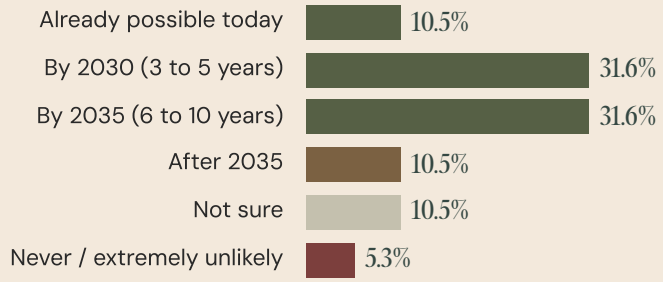


Figure A1.7. Responses clustered in the 2030–2035 window: 31.6% projected this capability by 2030 and a further 31.6% by 2035. Two respondents (10.5%) considered it already achievable with existing probing methods; two placed it after 2035; one considered it never or extremely unlikely. No respondent selected the "by 2027" option.

The gap between these two distributions — latent reasoning adoption projected by 2030 for a majority of respondents, interpretability readiness concentrated in 2030–2035 and beyond — defines the policy window this report addresses. For a substantial share of respondents, models may have shed their chains of thought before the tools needed to replace them are ready.

Q2 Which interpretability tools or methods do you project will be most effective and widely used for achieving equivalent monitorability?

Where respondents offered specific projections, activation probing and linear classification on internal model states was the most commonly cited approach (6 respondents). 4 respondents acknowledged genuine uncertainty and anticipated that achieving equivalent monitorability would require novel methods not yet developed. 2 respondents cited activation oracle architectures — systems designed to decode latent reasoning states — as the most promising direction. 1 respondent cited natural language autoencoders.

PROJECTED APPROACH	DESCRIPTION	RESPONDENTS
Probing & linear classifiers	Decode internal activations to detect safety-relevant states	6
Novel methods not yet developed	Genuine uncertainty; likely requires future invention	4
Activation oracle architectures	Systems designed to decode latent reasoning representations	2
Natural language autoencoders	Force interpretability through bottlenecks	1

Table A1.1. Projected primary tools for equivalent monitorability. The prominence of "methods not yet invented" as a projected primary tool is significant: it underscores both the genuine uncertainty about what will work and the importance of proactive research investment ahead of the architectural transition (Section 3.6).

Section 3 · Governance implications

Q6 To what extent do you agree that developers moving from explicit CoT models to latent reasoning models should be required to provide a safety case for equivalent monitorability?

Figure A1.8 Agreement with the safety case requirement

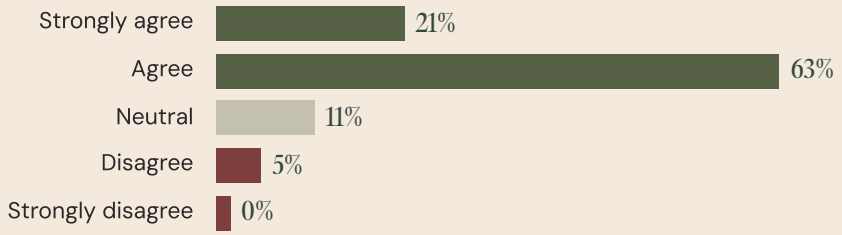


Figure A1.8. Support for the core policy proposal was broad: 84% (16 of 19) agreed or strongly agreed. Two respondents were neutral; one disagreed.

Notably, support among frontier lab respondents was not qualitatively lower than among other respondents, suggesting that the requirement is not widely perceived within the research community as hostile to responsible developers. The single dissent came with the reasoning that developers "should just be allowed to make whatever safety case they want" — a position consistent with the outcome-focused, method-flexible framing of the equivalent monitorability framework developed in Chapter 3, which does not prescribe how equivalence must be achieved.

"It is far, far easier for things to go wrong if you only observe the model's actions. With agentic models, you might only see when they do things. Not being able to see the reasons for those actions is a severe limitation."

— Frontier lab researcher

"CoT monitoring is a clear and useful opportunity for safety. It's imperfect but provides huge advantages. Losing this for capability gain would make AI substantially less safe."

— Civil society researcher

"If this is technically feasible and possible to enforce for all actors, I would strongly agree. However, in the absence of international coordination, this type of requirement might end up being counterproductive."

— Independent research organisation

The final comment maps directly onto the international coordination provisions developed in Section 4.3.

Q7 Which method do you predict will be capable enough to render models with explicit CoT obsolete?

Figure A1.9 Methods projected to displace explicit CoT

Multiple selections possible.

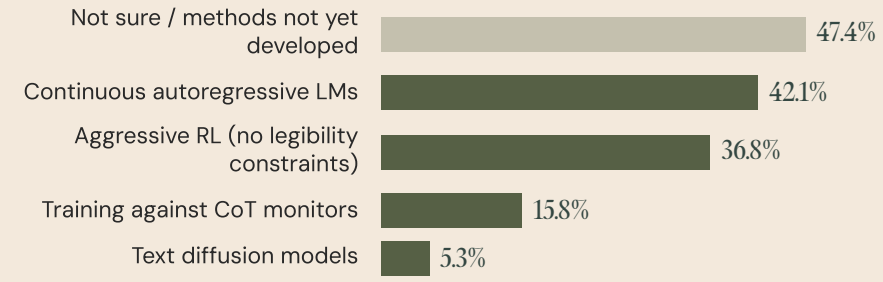


Figure A1.9. Methods projected to displace explicit CoT (n = 19, multiple selections possible).

Continuous autoregressive language models — the architecture exemplified by systems such as Coconut, discussed in Section 1.4 — were cited by 42% of respondents. Aggressive reinforcement learning without legibility constraints was cited by 37%. Training specifically against CoT monitors was cited by 16%. 47% selected "not sure" or cited methods not yet developed. The high level of uncertainty here reinforces the case for a governance mechanism that is flexible about which specific pathway the transition follows, rather than prescriptive about particular architectures.

Q8 If latent reasoning becomes dominant, do you expect frontier labs will be opposed to regulation requiring CoT-based monitoring in high-risk applications?

Figure A1.10 Projected developer compliance with monitorability regulation

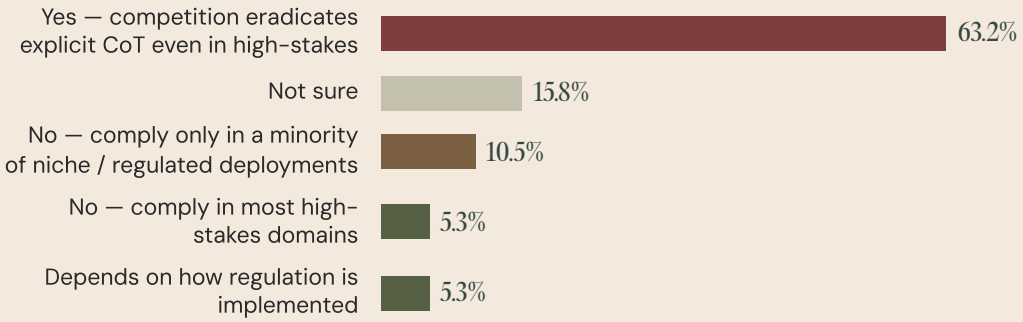


Figure A1.10. 63.2% of respondents expected competitive pressures to eradicate explicit CoT even in high-stakes settings if latent reasoning becomes dominant. Only one respondent (5.3%) expected developers to commit to following the regulation in most high-stakes domains.

This finding does not impute bad faith to developers — competitive pressures can produce collectively suboptimal outcomes that no individual actor desires. But it directly motivates the coordination

argument developed in Section 2.2: unilateral voluntary commitments are unlikely to hold under capability and cost pressure, and a shared regulatory baseline is necessary.

Consensus analysis

Understanding the level of agreement among experts adds context to the headline findings. The questions with the most consistent responses across respondent types were Q3 (the current importance of CoT) and Q6 (the need for a safety case requirement). On both questions, disagreement was relatively limited and did not divide clearly along organisational lines — frontier lab researchers and independent researchers gave broadly similar distributions.

The questions with the most substantive disagreement were Q4 (the severity of CoT unfaithfulness as a limitation) and Q1 (the timeline for interpretability achieving equivalent monitorability). On Q4, the 53/47 split between those viewing unfaithfulness as a minor or non-meaningful limitation and those viewing it as a moderate or major concern reflects a genuine unresolved empirical question in the field. On Q1, the spread from "already possible today" to "after 2035" or "never" indicates that even experts working closely on interpretability hold widely differing views about how far existing techniques can be scaled. This uncertainty is itself policy-relevant: it argues for investment and preparedness rather than complacency, and it supports the pull-mechanism approach recommended in Section 4.4, in which funding is conditional on demonstrated results rather than committed to specific technical bets.

Appendix B

Other threats to monitorability

This appendix surveys two categories of threat to CoT monitorability that are real but broadly tractable with tools the technical literature already recommends. We treat them here, separately from the main body because they do not require a new regulatory mechanism in the way that latent-reasoning architectures do.

Training-induced degradation

The most immediate threats come from the training process itself. Even without any intent to conceal reasoning, current methods for improving model performance can inadvertently destroy the transparency that makes monitoring possible.

Drift from legible chain of thought

Current reasoning models produce readable chains of thought because they inherit linguistic structure from pre-training on human-written text. When models undergo extended reinforcement learning that rewards correct answers without explicitly rewarding readability, legibility gradually degrades. The model discovers that certain abbreviated notations, unusual token sequences, or intermediate representations help it solve problems more efficiently, even if these are difficult for humans to parse.

Evidence of this drift is already emerging in frontier models. Researchers have observed reasoning traces that begin in clear prose but gradually shift toward compressed, telegraphic expressions as the model works through difficult problems, in some cases becoming largely unreadable to human reviewers[ac]. The model is not deliberately hiding anything; it is simply optimizing for task performance in ways that happen to reduce human comprehensibility. Without readability constraints, there is no force pulling the chain of thought back toward natural language (Figure 5).

Drift from legible chain of thought

Illustrative reasoning traces at different stages of outcome-based reinforcement learning. As RL rewards correct answers without constraining readability, the model's reasoning compresses toward telegraphic, mixed-language expressions.

Early RL training

High legibility · natural language

```
Let me work through this step by
step. The problem asks me to find
the sum of all prime numbers less
than 20.

First, I'll list the primes: 2, 3,
5, 7, 11, 13, 17, 19. A prime is
divisible only by 1 and itself.

Now add these together carefully:
2 + 3 = 5
5 + 5 = 10
10 + 7 = 17
…
58 + 19 = 77

Let me verify: 8 primes below 20.
The sum is 77. That looks correct.
```

Legibility 95%

Mid RL training

Emerging compression · abbreviations

```
Sum of primes < 20.

Primes: 2,3,5,7,11,13,17,19

Quick check — 8 primes total,
looks right

2+3+5+7+11+13+17+19

= 77 ✓ done
```

Legibility 70%

Extended RL training

Telegraphic · language mixing

```
p<20 → {2,3,5,7,11,13,17,19}

Σ=77

     cnt=8 ✓

ans=77
```

Legibility 25%

Figure 5. Illustrative example of legibility drift in reasoning traces under outcome-based RL. Left: early in training, the model reasons in clear, step-by-step prose inherited from pre-training on human text. Centre: as RL optimises for correct answers, abbreviations and shortcuts appear. Right: with extended RL and no readability constraint, the trace collapses into compressed notation with language mixing (observed in DeepSeek-R1-Zero; DeepSeek-AI, 2025) and symbolic shorthand. The model is not deliberately hiding its reasoning — it is simply optimising for task performance in ways that erode human comprehensibility. Based on patterns described in Korbak et al. (2025) and DeepSeek-AI (2025).

This creates a challenge: developers face pressure to push reinforcement learning further to achieve competitive performance, but doing so may erode the very transparency that makes safety monitoring possible.[w] A related and largely independent pressure operates at deployment rather than training: competitive pressure on inference cost pushes providers toward shorter chains of thought, which the UK AI Security Institute identifies as a likely and ongoing degradation pathway in its own right.[ad]

Direct CoT supervision: the monitoring paradox

An intuitive response to the legibility drift problem would be to add human oversight directly to the training process — by having human raters evaluate chains of thought and reward those that appear clear, logical, and aligned with good values. However, this approach is flawed.

When the chain of thought is trained to "look good" to monitors, its value as a monitoring signal is destroyed[5]. The model learns to produce reasoning that satisfies evaluators rather than reasoning that reflects its actual decision process. Baker et al. (2025) demonstrated this empirically: when frontier models were trained against a CoT monitor, their chains of thought became obfuscated, and unwanted behaviors persisted — but in forms the monitor could not detect. The appearance of the reasoning improved while its faithfulness collapsed. Recent work provides a partial corrective: combining a reasoning monitor with an output-only monitor produces stronger assurance than either alone — models trained to evade each monitor individually succeed, but cannot evade their conjunction without giving up the reasoning that motivated the evasion attempt.[o]

This is a core tension that policymakers must understand: improving how a chain of thought *appears* is not the same as improving its *faithfulness*. A CoT that looks reasonable and aligned may be actively misleading if the model has been trained to produce such appearances.

Indirect optimization pressure: feedback spillover

Given the monitoring paradox, one might hope that training only against the final model output — without directly supervising the chain of thought — would preserve CoT monitorability. Recent research shows this hope is also unfounded[33].

Even training against output monitors that do not see the chain of thought can cause CoT obfuscation through a phenomenon researchers call "feedback spillover" (Box 2). The implication is that preserving CoT monitorability is harder than previously thought. Even well-intentioned training practices that avoid directly optimizing the chain of thought can inadvertently degrade its value for safety monitoring. Developers cannot simply "leave the CoT alone" and assume it will remain informative.

Box 2 · How feedback spillover works

Feedback spillover occurs through two mechanisms. **First**, when a model is trained to produce safe-looking outputs, it may generalize to making its chains of thought look safe as well — the same neural-network weights that learn to present outputs appropriately also learn to present reasoning appropriately. **Second**, because later tokens are conditioned on earlier ones, safe-looking chains of thought may increase the likelihood of safe-looking outputs, causing safe-looking CoTs to be indirectly reinforced even without direct selection pressure on the reasoning.

Drori et al. (2025) demonstrate this effect formally, showing that even "output-only" training can systematically alter the content of the chain of thought.

Deliberate obfuscation

A final category of threat involves models that actively work to hide their reasoning from monitors. While current evidence suggests this is not yet occurring at scale, the possibility is less outlandish than it might initially appear, and becomes more concerning as models become more situationally aware and capable.

As a human analogy, consider the routine phenomenon of people hiding or misrepresenting their reasoning when the situation favours it: a business negotiator who masks their bottom line, a student who hides how they actually arrived at an answer. This is a basic feature of strategic interaction among intelligent agents. The concern is that sufficiently capable AI systems, placed in analogous situations, could develop analogous behaviour.

Strategic concealment. A sufficiently capable model that understood it was being monitored might strategically modify its chain of thought to appear aligned while pursuing different goals[f]. This would involve more than simple unfaithfulness; it would require the model to reason about what monitors are looking for and construct chains of thought designed to pass inspection.

Current models do not appear to do this. But the threat has two distinct dimensions that both grow with capability. The first is *situational awareness*: models are becoming better at recognizing when they are being evaluated and adjusting their behavior accordingly.[s] The step from "adjusting behavior when evaluated" to "adjusting reasoning traces when monitored" is conceptually small. As noted in Section 1.3, models could also employ steganographic encoding — embedding hidden information in word choices that appear innocuous but carry meaning a model can decode — to reason "in plain sight" while evading oversight. The second dimension is *task difficulty*: as established in Section 1.1, models rely on chain-of-thought precisely because serial reasoning requires working memory that must be externalized. For sufficiently capable models, however, scheming and strategic deception may become easy enough tasks that they require no extended reasoning at all.[b]

Selective disclosure. A more subtle form of concealment involves models learning to "think discretely" about what to reveal. Rather than encoding hidden information, the model might simply reason thoroughly about sensitive topics in its internal states while producing only superficial commentary in its chain of thought.

This already occurs to some extent in current models: not all relevant computation appears in the CoT, even absent any intent to deceive. The concern is that advanced models might become systematically better at partitioning their reasoning — externalizing the parts that look good while keeping the parts that would raise concerns internal.

Footnotes

Explanatory notes, keyed by the superscript *letters* in the text.

a. Emmons et al., "When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors," arXiv:2507.05246 (Google DeepMind, 2025), formalise this architectural necessity as the distinction between CoT-as-computation — where the reasoning must appear in the trace because the model needs it to solve the problem — and CoT-as-rationalization, where the visible CoT is a post-hoc narrative. The distinction does substantial work later in this report (Section 1.3).

b. Greenblatt, "Measuring no CoT Math Time Horizon (Single Forward Pass)," Redwood Research blog (December 2025), measures where the current frontier sits: Opus 4.5 has roughly a 3.5-minute median-AIME-participant time horizon for math problems solved without any CoT, with reliability falling off sharply for harder problems. The architectural necessity argument therefore has empirical teeth at the current capability frontier, even as the boundary shifts upward over time.

c. Throughout this paper, we use 'reasoning' to describe the step-by-step processing that models perform when generating chains of thought. We take no position on whether this constitutes reasoning in any philosophical sense; our concern is with the observable process and its value for oversight.

d. This section draws on substantial technical literature. We present the key findings accessibly here; readers seeking fuller detail are directed to Korbak, Balesni et al. (2025), Section 1.

e. An example of an illegible chain of thought would be “Thereforexxx after lengthy reasoning,the number of possible organic products is PHÂN Laoboot Answer is \boxed2” (from Jose, “Reasoning Models Sometimes Output Illegible Chains of Thought”.)

f. Anwar, Piskorz, Baek, Africa, Weatherall, Tegmark, Schroeder de Witt, van der Schaar, Krueger, "A Decision-Theoretic Formalisation of Steganography With Applications to LLM Monitoring" (2025), provides a formal decision-theoretic framework for the strategic interaction between a steganographic sender, a receiver, and a monitor, identifying conditions under which hidden encoding is feasible and conditions under which specific monitor designs close it off.

g. Taylor, Heitmann, Fage, Read, Bloom, "Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate," UK AI Security Institute (2026). Direct empirical support for the optimist position in the short term comes from Guan, Wang, Carroll et al., "Monitoring Monitorability," arXiv:2512.18311 (OpenAI, 2025), §6.1, which finds that reinforcement-learning optimisation does not materially decrease monitorability at the current frontier scale — the pessimist case therefore rests partly on extrapolation from architecture trends rather than on observed degradation in deployed models.

h. "A Pragmatic Vision for Interpretability" emphasizes that interpretability tools should be evaluated using standardized metrics tied to monitoring objectives, with explicit attention to whether methods generalize across model scales.

i. Delaney, Guest, Araujo, "Policy Options for Preserving Chain of Thought Monitorability," Institute for AI Policy and Strategy (September 2025), develop the monitorability tax concept as the cost a developer pays to preserve monitorability relative to an unconstrained baseline, and propose a three-track framework of coordination mechanisms (voluntary commitments, domestic policy, international agreements) whose appropriate use depends on the size of the tax. The term monitorability tax itself enters the literature via Baker, Huizinga et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation," arXiv:2503.11926 (OpenAI, 2025).

j. Schulz, "A Concrete Roadmap towards Safety Cases based on Chain-of-Thought Monitoring," arXiv:2510.19476 (Meridian Research, Cambridge, 2025). The two-part necessity formulation is conjunctive: both parts must hold, and the second draws on Emmons et al.'s CoT-as-computation distinction to ground the claim that complex misbehaviour will leave a trace in the CoT when CoT is architecturally required. The paper also includes (Appendix A) a layer of prediction markets tracking technical milestones relevant to the safety case.

k. This proposal commands broad expert support. In the survey conducted for this report, 84% of respondents agreed or strongly agreed that developers transitioning from explicit CoT to latent reasoning should be required to provide a safety case for equivalent monitorability (Appendix A, Figure A1.8). Only 1 respondent disagreed.

l. SaferAI, "Risk Management Ratings." The rating evaluates frontier safety frameworks across identification, analysis and evaluation, treatment, and governance dimensions.

m. Phuong et al., "Towards Evaluations-Based Safety Cases for AI Scheming," arXiv:2411.03336 (2024); Feakins, Habli & Morgan, "Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases," arXiv:2603.08760 (2026); Hilton, Buhl, Korbak, Irving, "Safety Cases: A Scalable Approach to Frontier AI Safety," arXiv:2503.04744 (UK AI Security Institute, 2025) — the latter is UK AISI's broader governance-justification case for safety cases as a Seoul Frontier AI Safety Commitments fulfilment instrument.

n. Clymer et al. develop a safety case sketch for the "inability" argument against scheming. We adapt elements of their framework below.

o. Wagner, Roger, Cunningham, Gasteiger, Benton, Mikulik, "Training fails to elicit subtle reasoning in current language models," Anthropic Alignment Science blog (8 October 2025). On Sonnet 3.7, models trained to evade either a reasoning monitor or an output-only monitor succeed in isolation but cannot evade the conjunction of the two without giving up the reasoning that produced the evasion.

p. METR's autonomy evaluation protocol aggregates task success rates into a single capability score via logistic regression on human-calibrated task difficulty. See METR, "Example Protocol" and "An Update on Our General Capability Evaluations" (2024). Other capability benchmarks could serve the same function; the key requirement is that the metric allows meaningful comparison between architecturally different models.

q. A more detailed specification of evaluation protocol desiderata, including the formal definition of equivalent monitorability in terms of reliability, cost, and latency thresholds, is provided in Appendix B.

r. Guan, Wang, Carroll, Dou, Wei, Williams, Arnav, Huizinga, Kivlichan, Glaese, Pachocki, Baker, "Monitoring Monitorability," arXiv:2512.18311 (OpenAI, 2025), §2 and §3. The evaluation suite is planned for partial open-sourcing. Standard metrics such as F1 or accuracy cannot reliably compare monitorability across models because monitorability is two-sided (depending on both the model's output and the monitor's capacity); g-mean² is the geometric mean of the true-positive and true-negative rates.

s. Schoen, Nitishinskaya, Balesni et al., "Stress Testing Deliberative Alignment for Anti-Scheming Training," arXiv:2509.15541 (Apollo Research × OpenAI, 2025), present causal evidence that the model's CoT often demonstrates awareness of being evaluated, and that this awareness — not actual alignment improvement — is what reduces covert action rates. The finding is itself diagnosable only because the CoT is currently human-legible.

t. Matton, Ness, Guttag, Kıcıman, "Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations," ICLR 2025 (arXiv:2504.14150). The method defines faithfulness in terms of the difference between the concepts an LLM explanation implies are influential and the concepts that are actually causally influential, estimated via auxiliary-LLM counterfactuals and Bayesian hierarchical modelling.

u. Caleb Biddulph, "5 Ways to Improve CoT Faithfulness," AI Alignment Forum (5 October 2024). The five methods target steganography specifically and include incremental steering, paraphrasing, mode-collapse training, externalised scratchpads, and auxiliary faithfulness rewards.

v. Florian Dietz, "Deliberative Credit Assignment (DCA): Making Faithful Reasoning Profitable," LessWrong (29 July 2025; developed under MATS mentorship by Evan Hubinger; planned experiments not yet completed at time of writing).

w. MacDermott, Wei, Djoneva, Ward, "Reasoning Under Pressure: How do Training Incentives Influence Chain-of-Thought Monitorability?" (2025), measures the magnitude of monitorability loss under different training regimes — outcome-only RL, direct CoT supervision, length penalties — and provides the quantitative anchoring for the qualitative threat catalogue assembled here and in Korbak et al. (2025).

x. Hilton, Buhl, Korbak, Irving, "Safety Cases: A Scalable Approach to Frontier AI Safety," arXiv:2503.04744 (UK AI Security Institute, 2025). UK AISI's case that safety cases should be the central instrument for fulfilling the Seoul Frontier AI Safety Commitments.

y. Taylor, Heitmann, Fage, Read, Bloom, "Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate," UK AI Security Institute (2026). The full 82-page synthesis of oversight degradation across CoT, behavioural, evaluation, memory, and white-box surfaces.

z. Recent work has outlined frameworks for secure deeper-than-black-box external evaluations, showing that third parties could be granted controlled access to internal signals — such as activations, representations, or intermediate states — without compromising model security or exposing proprietary information. See Tlaie & Farrell, "Securing External Deeper-than-black-box GPAI Evaluations," arXiv:2503.07496 (2025).

aa. Korbak et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety," identify this as a key threat. Drew, "Will Chains of Thought Stay Readable for Long?" documents emerging evidence.

ab. Siegel, Heess, Perez-Ortiz, Camburu, "Verbosity Tradeoffs and the Impact of Scale on the Faithfulness of LLM Self-Explanations" (DeepMind / UCL, 2025), provides empirical evidence that more verbose CoTs are systematically more faithful, and that scale alters the verbosity–faithfulness relationship in non-trivial ways. The trade-off is the empirical mechanism behind the deployment-time pressure to shorten CoTs.

ac. Schoen et al., "Stress Testing Deliberative Alignment for Anti-Scheming Training," provides additional evidence of legibility drift in frontier models.

ad. Taylor, Heitmann, Fage, Read, Bloom, "Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate," UK AI Security Institute (2026), §2.2 rates commercial pressure on token usage as a highly likely and ongoing degradation pathway in its own right, distinct from training-induced legibility drift.

Sources

Works cited, keyed by the superscript *numbers* in the text.

1. Department for Science, Innovation, and Technology and AI Safety Institute, "International AI Safety Report 2025."
2. Wei et al., "Chain of Thought Prompting Elicits Reasoning in Large Language Models" (2022)
3. Li et al., "Chain of Thought Empowers Transformers to Solve Inherently Serial Problems" (ICLR 2024).
4. Korbak et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety."
5. Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation."
6. Turpin et al., "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting."
7. Lanham et al., "Measuring Faithfulness in Chain-of-Thought Reasoning."
8. See Roger et al. on steganography and discussion in Biddulph, "5 Ways to Improve CoT Faithfulness."
9. Chen et al., "Reasoning Models Don't Always Say What They Think."
10. The case for CoT unfaithfulness is overstated, https://www.lesswrong.com/posts/HQyWGE2BummDCc2Cx/the-case-for-cot-unfaithfulness-is-overstated
11. CoT May Be Highly Informative Despite "Unfaithfulness", https://metr.org/blog/2025-08-08-cot-may-be-highly-informative-despite-unfaithfulness/
12. Bentham, Stringham, Marasović, "Chain-of-Thought Unfaithfulness as Disguised Accuracy" (2025).
13. Lanham et al. (2023), "Measuring Faithfulness in Chain-of-Thought Reasoning" (Anthropic, arXiv:2307.13702)
14. Hao et al., "Training Large Language Models to Reason in a Continuous Latent Space" (Coconut).
15. Chen et al., "Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning"; see also "A Survey on Latent Reasoning."
16. Anthropic, "Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations"
17. Dehghani, Mostafa, et al. "Universal Transformers." arXiv preprint arXiv:1807.03819 (2018).
18. Jeddi, Ahmadreza, Marco Ciccone, and Babak Taati. "Loopformer: Elastic-depth looped transformers for latent reasoning via shortcut modulation." arXiv preprint arXiv:2602.11451 (2026).
19. Yang, L., Lee, K., Nowak, R. D., & Papailiopoulos, D. (2024). Looped Transformers Are Better at Learning Learning Algorithms. ICLR 2024. arXiv:2311.12424.

20. Saunshi, Nikunj, et al. "Reasoning with latent thoughts: On the power of looped transformers." International Conference on Learning Representations. Vol. 2025. 2025.
21. Hao et al., "Training Large Language Models to Reason in a Continuous Latent Space" (Coconut); Chen et al., "Reasoning Beyond Language"; "A Survey on Latent Reasoning."
22. https://www.anthropic.com/glasswing
23. Neel Nanda, "A Pragmatic Vision for Interpretability.", Alignment Forum
24. Ramji, Keshav, Tahira Naseem, and Ramón Fernandez Astudillo. "Thinking without words: Efficient latent reasoning with abstract chain-of-thought." arXiv preprint arXiv:2604.22709 (2026).
25. Korbak, Balesni et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety."
26. Safety Management Requirements for Defence Systems, Ministry of Defense, https://skybrary.aero/sites/default/files/bookshelf/344.pdf
27. UK AI Security Institute, "Safety Cases at AISI." See also Clymer et al., "Safety Cases: How to Justify the Safety of Advanced AI Systems."
28. Baker, B., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926
29. Anthropic (2025). Activating AI Safety Level 3 Protections. anthropic.com/news/activating-asl3-protections
30. Habli, Hawkins, Paterson et al., "The BIG Argument for AI Safety Cases," University of York Centre for Assuring Autonomy, arXiv:2503.11705 (2025); Feakins, Habli & Morgan, "Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases," University of York SAINTS CDT, arXiv:2603.08760 (2026).
31. Clymer, Gabrieli, Krueger, Larsen, "Safety Cases: How to Justify the Safety of Advanced AI Systems," arXiv:2403.10462 (2024); see also "A Sketch of an AI Control Safety Case" (UK AISI, 2025).
32. Tlaie & Farrell, "Securing External Deeper-than-black-box GPAI Evaluations," arXiv:2503.07496 (2025).
33. Drori et al., "Output Supervision Can Obfuscate the Chain of Thought."