BOARD BRIEFING . SEPTEMBER 2026

Frontier AI: A *Board-Level Question* Now

What harness engineering means, and the four questions Boards should be asking.

Since April 2026, when Anthropic announced Claude Mythos Preview and declined to give it public release, citing the security risk posed by its own offensive cybersecurity capability, Boards across sectors have been asking the same question in different words: what exactly are we doing about frontier AI. The question has moved out of the technology function and onto the Board agenda, and it deserves a considered answer rather than a reflexive one.

A recent article from the Bank of England lays out the terrain with unusual precision. It sets aside the marketing language that surrounds most AI commentary and walks through the practical options open to firms deploying frontier models, along with the tradeoffs each option carries.

A term worth learning

Harness engineering is a term that only entered common use in 2026. It describes everything built around a frontier AI model: the tools it can call, the data it can see, the checks applied to its output, and the workflow that routes its findings to a human. The term is deliberately broader than "control environment." A control environment, in the traditional risk sense, covers oversight and governance. A harness covers that and more: the engineering choices that determine whether a powerful model produces something useful at all.

For Boards, this distinction matters. A firm can have excellent AI governance policy and still deploy a model that produces unusable output, because nobody engineered the surrounding capability properly. The reverse is equally possible: a technically strong harness with no governance behind it. Both are live risks, and neither is solved by the other.

Four questions Boards should be asking

- **Capability.** Model access on its own does not create a working defence tool. The output only becomes useful once someone has designed how tasks are framed, how results are checked, and how findings reach the right team. Firms have three broad paths open to them: build the harness internally, buy a vendor platform, or assemble it from open-source components. Each carries a different cost. Vendor platforms are quick to deploy but weaker on customisation. Internal builds are flexible but consume engineering time that could go elsewhere. Open-source components move fast, sometimes faster than internal teams can track, and few in-house teams have evaluated more than one or two of them. Increasingly, firms are adding an orchestration layer on top, so that different tools handle discovery, validation, and remediation routing separately, rather than betting everything on one vendor or one model.
- **Governance.** The central question here is how much of the organisation a model is allowed to see. Source code, architecture diagrams, asset inventories, and control data all improve a model's usefulness, and all raise the stakes if the model or its harness is compromised. Governance has to answer this in specific terms: which repositories are in scope, which are excluded, what supplier assurance is required before a third-party harness touches anything production-adjacent, and whether the model runs in a sandbox, a production-like test environment, or live infrastructure. Vague answers here tend to default to whatever the loudest voice in the room prefers, which is not a governance position, it is an absence of one.
- **Control environment.** This is where governance intent becomes enforcement. Network segregation, access permissioning, approval workflows, monitoring, and defined rules of engagement all sit at this layer. The guardrails supplied by the model vendor are a starting point, not a finishing line: additional controls have to be built into the harness itself and

into the environment it operates in. Firms that treat vendor guardrails as sufficient are, in effect, outsourcing a decision that belongs to them.

- **Operations.** None of the above matters if findings pile up faster than teams can act on them. Scaling a frontier AI capability depends on the capacity to triage, validate, and remediate at volume, without burying engineering teams in false positives. This is a people and process problem as much as a technology one, and it requires a blend of red-teaming skill, software engineering, security architecture, and AI literacy that most organisations have not yet assembled under one roof.

The Bank of England article deliberately avoids prescribing a single implementation model, and it says so directly. That is a reasonable choice given how quickly the underlying tools are changing, but it also leaves firms to interpret the four questions above according to their own size, complexity, and risk appetite. The article is written for financial services, where the Bank of England has direct authority, but the same four questions apply to any sector where competitive pressure is pushing firms to deploy frontier AI quickly. Financial services is simply where the framework surfaced first.

Our View

At BeatQuantum, we treat AI security as a distinct discipline from conventional cybersecurity, not a variant of it. In a typical cyber scenario, an uninvited attacker is trying to breach a perimeter you built to keep them out. With frontier AI, the situation reverses: you are inviting something powerful in, giving it access to sensitive data and live processes, on the assumption that it will behave. That is a different threat model, and it calls for a different security mindset.

Picture a frontier AI agent as a croupier in a casino. It has access to the money, it deals the cards, and it knows exactly where every camera is pointed. A casino with corrupt or careless croupiers can lose millions before anyone notices, no external thief required. That is why harness engineering exists: it is the set of controls, checks, and dealer rotations that keep a croupier honest even when nobody is watching every hand.

Extend the analogy one step further. A croupier trained at one casino can walk into a rival casino, immediately spot which of their croupiers are inexperienced, and exploit that weakness for profit. That is the shape of a frontier AI driven attack: capability trained in one context, redirected against a target that has not built equivalent defences.

We track regulatory guidance across major jurisdictions and help firms design technology resilience programmes built for this threat model. We developed two concepts specifically to counter frontier AI driven attacks: *Zero Trust Zero Tolerance*, which extends zero-trust principles to remove any implicit trust granted to an AI agent's actions, and *White Box Purple Test*, which combines purple-team testing methodology with full visibility into how the model's

harness is constructed. Both exist because the old assumption, that you are defending against an outsider, no longer holds once the thing with access to your systems is one you invited in yourself.

Full methodology: beatquantum.com/frameworks

Sources

1. Bank of England. “Frontier AI: Harness engineering.” Frontier AI Information Sharing Forum, 2 September 2026. bankofengland.co.uk/research/fintech/frontier-ai-information-sharing-forum/frontier-ai-harness-engineering

2. Böckeler, Birgitta. “Harness engineering for coding agent users.” martinfowler.com/articles/harness-engineering.html