



TOPICS:

Technology

SOURCE:

Bank of England

BoE: Frontier AI - Harness Engineering

- The article identifies the **AI harness** - the tools, workflows, controls, data and operating environments surrounding an AI model - as a critical determinant of whether frontier AI capabilities can be translated into usable cyber-defence outcomes. Effective harnesses should provide appropriate organisational context, facilitate validation and human review, control access to tools, and integrate findings into existing cybersecurity and engineering processes.
- Firms may combine **internally developed, vendor-provided, managed third-party and open-source components**. Each involves trade-offs between flexibility, transparency, supportability, customisation and engineering requirements. It is highlighted modular orchestration or wrapper layers as a means of coordinating specialised tools and models while reducing dependence on a single technology or supplier.
- A central risk-management consideration is balancing AI access to organisational context against **information sensitivity**. Source code, security architecture and asset information can improve model effectiveness but create confidentiality, data-protection, supplier and misuse risks. Isolated, sandboxed or production-like

- environments may therefore provide benefits while limiting direct interaction with live production systems.
- Controls should be embedded throughout the harness and operating environment, including **access restrictions, network isolation, monitoring, environment segregation, approval workflows and rules of engagement**. Model-level guardrails alone may be insufficient.
 - Finally, successful scaling depends on more than vulnerability detection. Firms require adequate capacity to **validate, triage, prioritise and remediate findings**, particularly given false positives and potential remediation burdens. Effective adoption consequently depends on operating models combining AI expertise with red teaming, software engineering, security architecture and operational cyber capabilities. The overall message is that frontier AI's cyber-defence value depends substantially on the surrounding **architecture, controls, people and processes**, rather than model capability alone.