

UK AI Security Institute sets out the technical practices it wants frontier AI developers to adopt

Source copy kept by wheresthe.ai for the story "UK AI Security Institute sets out the technical practices it wants frontier AI developers to adopt" (<https://wheresthe.ai/d/gb-aisi-sets-out-priorities-for-technical-best-practice-in-frontier-ai-2026-10/>).

Source: <https://www.aisi.gov.uk/blog/towards-technical-best-practice-for-frontier-ai-development>

Text saved 25 Oct 2026 (the page as published).

SHA-256 of the saved text: 9e98380f4270e151ff9b368da6441f9df6fe7fa7259af76197708271671ad1f1

Documents the source offers: <https://www.ncsc.gov.uk/sites/default/files/documents/Guidelines-for-secure-AI-system-development.pdf>

On 22 September, the Prime Minister set out steps the UK is taking to ensure AI delivers on its vast potential for good. These include a commitment to work towards principles and standards that ensure the most capable AI systems are developed securely and with appropriate transparency. The AI Security Institute (AISI) will play a key role in delivering that commitment, with partners across industry, governments and beyond. This work responds to continued, rapid AI capability progress in recent months. That progress promises a wide range of benefits, but it has also outpaced several practices used to manage risk and oversee frontier systems – especially with risks now emerging earlier, during AI development itself. Here we set out AISI's view of these challenges and the new technical practices and controls most urgently needed to address them. We draw on three years of experience testing, evaluating and researching frontier AI models, in partnership with industry and other technical authorities such as the Center for Advancing Innovation and Standards for Super Intelligence. We also explain how we aim to work with developers and others to help establish best practice that could later inform formal technical standards. AISI is not seeking to mandate requirements for frontier developers, but to contribute our expertise to an industry-wide effort to advance technical standards and best practice. We welcome leading developers' recent commitments to that effort, including through the White House Accord on Super Intelligence. How voluntary action already shapes AI development Much of how frontier developers manage risk today is voluntary. Developers publish safety frameworks, share models with third parties for pre-release evaluation and, most recently, have begun embedding third-party evaluators. AISI has helped establish these practices. Working with industry, we have independently evaluated more than 55 models before release, developed principles for safeguards evaluation, helped shape risk assessments and safety frameworks, and jointly tested agent control monitors. Beyond industry collaboration, we have released open-source evaluation tools including Inspect – now the standard framework for leading third-party evaluators – and published field-building research on topics such as safety cases and monitorability and control. We have advanced the science of evaluation internationally by coordinating the Network for Advanced AI Measurement, Evaluation and Science and contributing to a NIST taxonomy. Elsewhere in government, experts at the NCSC, DSIT and DCMS have worked with industry and international partners, including the US and other governments, to develop a code of practice on AI cyber security which now forms the basis of ETSI EN 304 223. Together, these practices have helped industry operate more safely and given governments and the public a clearer view of how AI capabilities and risks have progressed. The work we set out today builds on that foundation. Where new challenges have emerged Since these practices were established, AI capabilities have continued to advance rapidly. Through our technical work, we have identified two new challenges facing frontier AI organisations today. First, risks are arising earlier in the development process. Evaluations and other security practices have largely centred on making models safe for public release. But today's AI systems are autonomous and capable enough, particularly in cyber, to pose risk even while unreleased: during training, internal deployment, and evaluation itself. Recent incidents demonstrate how inadequate control and monitoring at these stages could lead to real harm. Second, evaluation practices are struggling to keep pace. Capability benchmarks are often saturated soon after they're created, and building credible new evaluations takes time: they now entail increasingly complex tasks, multi-agent settings, multi-day runs, and tighter security controls – all while accounting for the effects of inference compute scaling, tooling and scaffolding. Evaluation methodologies for alignment and control are still nascent, meaning they can't assure whether increasingly autonomous AI agents will behave as intended, or whether controls will reliably catch any misbehaviour. Several trends exacerbate these challenges. As AI systems become more capable, they may become harder to oversee. Models' growing awareness of when they're being tested may allow them to mask how they would behave when deployed. Multi-agent settings may trigger new behaviours that have barely been studied. And, as AI contributes more to its own development, progress could accelerate, meaning these problems

may arise faster and with less direct human oversight involved. It's these challenges that call for new technical best practice, to enable frontier AI development activities to continue safely as capabilities progress. Our view on priorities for new best practice Below we set out our view on the most urgent priorities for new technical best practice. The list is not exhaustive, and we aren't suggesting AISI should address every item. Many priorities are interlinked, and align with practices to which companies have publicly committed: Assessing and documenting risk: how should frontier AI organisations assess risk and report on it throughout development; what should risk assessments cover; and what standard of evidence and documentation should they meet. Overseeing AI systems throughout their lifecycle: what controls, monitoring, detection, and other mechanisms should be used to ensure AI systems operate as intended during training, evaluation and internal and external deployment; how should frontier AI organisations share best practice; and how should existing guidelines for secure AI development be implemented and kept up to date as capabilities evolve. Preserving and evaluating oversight mechanisms: what technical choices will preserve developers' ability to oversee AI systems as they become more capable; and how should oversight mechanisms be rigorously evaluated. Detecting, responding to, and reporting agentic AI incidents: how should unsafe agent behaviour be defined and its severity determined; what technical systems are needed for rapid detection and response; and what principles will govern how incidents are reported to affected parties, governments, and the public, aligning with existing standards . Sharing information on the speed of AI development: which metrics best track progress, including AI's use in its own development as an indicator of recursive self-improvement; and when should these metrics be shared with governments and the public, so that rapid progress (or stagnation) does not come as a surprise. Enabling independent auditors and evaluators: what expertise qualifies auditors and evaluators to make effective safety assessments; what scope should frontier AI organisations grant to embedded auditors and other independent third parties; what access and resources should they be given; and what should they report to companies, governments and the public. Measuring capabilities in security-relevant domains: how can internal and external evaluators measure the full extent of system capabilities; and which practices in elicitation, environmental realism, and evaluation design give the most accurate results. Measuring AI alignment: which techniques robustly evaluate how a model behaves in various high-stakes situations; and what are the limits of what these evaluations show. Technical practices will need to evolve as rapidly as capabilities do. What works for today's systems may not hold for more capable ones, given the trends we identified earlier. Best practice will therefore need to be developed with future models in mind, kept under continual review, and updated as new capabilities arrive. The path to technical best practice An agenda so broad and complex is best delivered through collaboration between frontier developers and other technical organisations. It calls for a wide range of expertise, including in cyber security, incident response, national security and safety-critical industries with deep experience of managing risk in complex systems. Working together makes effort additive rather than duplicative, and lets each organisation's strengths cover others' blind spots. In our view, the first step is to meet the immediate technical need, with industry and other frontier AI organisations iteratively developing and sharing practices that address the challenges developers face today. As practices mature, and where consistency and interoperability add value, they could be formalised via international standards development organisations – establishing common security baselines that let developers around the world operate with confidence. AISI will contribute its technical expertise, built on years of experience in evaluation, research and technical collaboration. We will work with frontier developers and others to develop voluntary practices, leading where valuable, and supporting others' efforts where we can. If you would like to collaborate on any of these problems, we'd like to hear from you.