Key Takeaways on the Challenges in Frontier AI Governance

22.5.2026 - Louise Marie Hurel | The Royal United Services Institute for Defence and Security Studies

The RUSI Cyber and Tech research group held an event at RUSI Europe on 12 May focusing on a central challenge in frontier AI governance, exploring what constitutes 'sufficient and appropriate access' for external evaluators, regulators and researchers.

Current approaches remain fragmented and largely dependent on private companies, while there is still no shared operational definition of access, risk, or appropriate safeguards. The initiative brought together around 40 experts from cybersecurity, frontier AI labs, evaluation organizations, and policy institutions to develop a baseline framework for secure third-party access to frontier models.

Participants examined different categories and levels of access, as well as the risks associated with them. A major challenge is the difficulty of distinguishing between existing and potential threats in a rapidly evolving technological environment. To address this, the group developed an access-risk matrix assessing the relationship between access levels, structural exploitation risks, and cybersecurity concerns. Risks such as credential compromise, leakage of commercial information, and malicious misuse were classified as low, medium, or high, alongside mitigation measures including role-based permissions, segregation of duties, monitoring systems, and controlled evaluation environments.

The discussion emphasised that what was previously considered sufficient access is no longer adequate for advanced frontier models. Earlier evaluations based mainly on API access are increasingly limited because models have become more capable and adaptive. Access therefore needs to deepen while remaining proportionate to evaluation goals and associated risks. Participants also raised concerns about 'evaluation awareness', where AI systems recognize that they are being tested and modify their behaviour accordingly. This undermines the reliability of evaluations and may require deeper forms of access, including access to raw chain-of-thought reasoning and latent representations within models, in order to conduct more representative assessments.

At the same time, participants stressed that external access creates major security concerns, including cybersecurity vulnerabilities, trade secret exposure, and the risk of malicious misuse. The challenge is therefore to enable meaningful evaluation without compromising model security.

Several broader governance challenges emerged, including the absence of shared definitions, the rapid evolution of AI capabilities and questions around legitimacy and international coordination. Participants highlighted the importance of harmonised terminology, shared standards, feedback mechanisms, and broader international participation beyond Western institutions.

Overall, the discussion concluded that the priority is not yet a definitive regulatory model, but the creation of a shared framework capable of balancing evaluation needs, security concerns, and international cooperation in an evolving AI landscape. Talks about an evaluation company or an AI institute were also underlined.

https://www.rusi.org/event-summaries/key-takeaways-challenges-frontier-ai-governance