Research Papers

May 2026

Developing a Framework for Secure Third-Party Access to Frontier AI

Secure Access to Frontier AI Taskforce (SAFA-TF) Report

Louise Marie Hurel, Elijah Glantz and Daniel Cuthbert

RUSI

Disclaimer

The content in this publication is provided for general information only. It is not intended to amount to advice on which you should rely. You must obtain professional or specialist advice before taking, or refraining from, any action based on the content in this publication.

The views expressed in this publication are those of the authors, and do not necessarily reflect the views of RUSI or any other institution.

To the fullest extent permitted by law, RUSI shall not be liable for any loss or damage of any nature whether foreseeable or unforeseeable (including, without limitation, in defamation) arising from or in connection with the reproduction, reliance on or use of the publication or any of the information contained in the publication by you or any third party. References to RUSI include its directors, trustees and employees.

RUSI Research Papers, May 2026

ISSN 2977-960X

Publications Team

Editorial

Director of Publications: Alice Trouncer

Managing Editor: Sarah Hudson

Assistant Editor: Sophie Boulter

Design

Graphic Designer: Lisa Westthorp

Research Editorial

Head of Research Governance and Editorial: Elias Forneris

Cover image: Panther Global Media / Alamy

Get in touch

www.rusi.org

enquiries@rusi.org

+44 (0)207 747 2600

The Royal United Services Institute for Defence and Security

61 Whitehall, London

SW1A 2ET

United Kingdom

Follow us on

Contents

Disclaimer

This project was made possible with the support of Google. The views expressed in this paper do not necessarily represent the views or policies of Google.

Methodology

This paper principally draws on findings from three workshops organised under the RUSI Secure Access to Frontier AI Taskforce (SAFA-TF), and consultations with individual experts. Overall, the paper reflects engagement with 39 experts. Held virtually from October to December 2025, the workshops brought together leading international AI engineers, evaluators, red teamers, cybersecurity specialists and policy and governance experts from civil society, government and the private sector (including frontier labs). The workshops were divided into three themes: conceptualising third-party evaluation risk, charting evaluation and access-type risks, and identifying pathways towards mitigating risk. The Access–Risk Matrix and the mitigation strategies outlined in this paper were developed throughout the course of the workshops, and through requests for input on earlier versions of both in between the sessions. In January and February 2026, SAFA-TF participants and additional subject-matter experts peer-reviewed previous versions of this paper and provided comments and feedback ahead of publication.

The scope of this paper, and its accompanying SAFA-TF workshops, is specifically limited to risks and mitigation measures of closed frontier AI models and systems in pre- and post-deployment contexts. Open source models and models with niche or discrete applications were not considered in the research. Likewise, the paper and workshops do not centre on specific evaluation methodologies, but rather on the broader ecosystem in which the range of tests is conducted. The evaluation and access types considered are strictly by third-party evaluation companies or independent researchers and not by government regulators or law enforcement. The authors acknowledge that the risks, severity levels and controls are not exhaustive. The existing gap between risks that have been publicly identified and potential risks is also worth bearing in mind, as it means the Access–Risk Matrix is a living, evolving contribution to the debate. Moreover, as identified during the discussions, the authors believe (and highlight in the final chapter) that further dialogues and collaboration between evaluators and cybersecurity researchers are needed to ensure benchmarking methodologies are ever more attuned to the changing threat landscape and the tactics, techniques and procedures of adversaries.

Note that an AI language model was used to support early literature review and proofreading, and provided suggestions on how to illustrate the data collected for the ‘Exploitation Pathways’ section, the Access–Risk Matrix in an image and table, and in the development of the Key Terms and Concepts in a language that is easily accessible to broader non-expert audiences.

Key Terms and Concepts

Frontier AI model: A general-purpose AI model – typically a large language model (LLM) – that represents the cutting edge of capability, often trained at a significant computational scale. Examples include GPT-5.2, Claude Opus 4.6 and Gemini 3.1 Pro. These models may pose novel safety and security risks due to their advanced capabilities and varied applications.

AI model versus AI system: A *model* is the trained artefact – the neural network with its architecture and weights. A *system* is what results when that model is integrated into a specific application with a user interface, data pipelines, safety filters and a deployment context. The same model can underpin many different systems with different risk profiles. The EU AI Act regulates both, but through distinct mechanisms.

Closed versus open models: Closed (or proprietary) models are accessible only through controlled interfaces such as APIs (application programming interfaces) or chat products, and the model weights are not publicly available. Open (or open weights) models make their weights available for download, allowing anyone to run, fine-tune or inspect the model.

Third-party evaluation (of a model): An assessment of a frontier AI model's capabilities, safety properties or security posture, conducted by an organisation distinct from the model's developer. These evaluations include dangerous capability evaluations, adversarial testing (red teaming) and robustness assessments. This assessment is distinct from conformity assessments of deployed AI systems, which evaluate compliance with sector-specific requirements.

Access levels (black-box, grey-box and white-box access): The degree to which an external evaluator can interact with or inspect a model. These range from *black-box access* (querying the model and observing outputs only) through *grey-box access* (limited visibility into internal components such as logits or activations) to *white-box access* (full access to model weights, architecture, training data and development documentation). Different evaluation types require different levels of access, each carrying distinct security implications.

Pre-deployment versus post-deployment evaluation: Pre-deployment evaluations are conducted before a model is released – typically on development checkpoints – to identify dangerous capabilities or safety gaps before they reach users. Post-deployment evaluations assess models already in use, monitoring for emergent risks or verifying that safety mitigations hold. As this paper shows, pre- and post-deployment differ in access conditions, security risks and evaluator relationships.

Acknowledgements

The authors would like to thank all those who volunteered their time and expertise as part of the Secure Access to Frontier AI Taskforce (SAFA-TF). In particular, the authors would like to acknowledge the contributions of all SAFA-TF members and participants throughout the workshops and their feedback throughout the development of this paper: Daniel Cuthbert, Mohamed Samy, Ehab Hussein, Omer Nevo, Alejandro Ortega, Kevin Klyman, Markus Anderljung, George Balston, Francesca Federico, Charles Foster, Esme Harrington, Pia Huesch, Kellin Pelrine, Adriana Stephan, Raquel Vazquez, Pegah Maham, Talita Dias, Rumman Chowdhury, Robert Trager, Conrad Stosz, Dawn Song, Abby Cruz, Mathias Vermeulen, Clément Briens, Markus Hobbhahn, Madeline Carr, Ingrid Dickinson and others who provided comments to earlier versions of this paper.

Executive Summary

As frontier AI models expand in their capability and application, their evolution must remain grounded in safety and security safeguards.

Third-party evaluation of frontier AI models is increasingly recognised as essential to supporting their safety and security by developers, governments and regulators alike. Yet, enabling meaningful external evaluation requires granting access to some of the most sensitive intellectual property in the tech/AI sector. The security risks associated with this access – from intellectual property (IP) leakage to model compromise to exploitation by state-sponsored actors – remain poorly mapped and inadequately standardised. This gap stifles the evaluation ecosystem, making it one where developers restrict access out of security concerns, while evaluators lack the information they need to conduct effective assessments.

Drawing on the work of the SAFA-TF, this paper proposes a shared framework for understanding and managing these risks. The aim is to move the conversation beyond the current tension between openness and security, or security and innovation, towards a practical, shared understanding of how to enable secure and effective third-party evaluation at scale.

Core Contributions

1. This paper develops a threat taxonomy of seven security risk categories specific to the context of third-party evaluation: Model Theft, Capability Reconstruction, Model Manipulation, Jailbreak/Safety Bypass, Accidental Exposure, Credential Compromise and Access Persistence, with Weaponisation as a potential resultant outcome. Each category is defined, illustrated with real-world examples and research evidence, and situated within a risk hierarchy organised by access depth and actor sophistication.
2. The paper proposes an Access–Risk Matrix that maps six types of evaluator access ranging from query-level inference to model internals, training data, evaluation data and compute infrastructure, against each risk category. The matrix assigns indicative severity ratings for each access-related risk, thus providing a structured basis for identifying the best-suited security mitigations based on the threat model.

3. Building on the Access–Risk Matrix, the paper proposes security mitigations organised by the previously identified security risk categories. Mitigations are distinguished as technical, procedural or contractual measures, and responsibility is assigned to developers, evaluators or both. These controls are grounded in established security principles (for example, least privilege, assume breach, need-to-know, data minimisation, time-bound access and proportionality) adapted for the specific context of frontier AI model evaluation.
4. The paper proposes a shared governance framework for maturing the third-party evaluation ecosystem and presents it as three pillars of action for the multistakeholder community:
    - Harmonise language and access tiers across the ecosystem.
    - Operationalise secure access through shared standards and practices.
    - Build feedback loops that allow the framework to evolve as the threat landscape, regulatory requirements and evaluation methodologies develop.

Introduction

Creating a Shared Understanding of How to Secure Access to Frontier AI

Safety and security assessments of frontier AI models should be seen as integral and iterative components of responsible AI development. A clearer, shared and cross-stakeholder understanding of the security risks associated with the different levels of model access could help advance the discussions beyond the tensions outlined above. In practice, this means enabling evaluators to conduct meaningful assessments while giving developers the confidence that internal and external measures are in place to ensure the security and highest quality of assessments.

Establishing common terminology, a shared understanding of risks and best practices for secure evaluation access is, as this paper argues, paramount to both enabling independent assessments at scale and managing security, reputational and societal risks.

This paper seeks to contribute to a common language and establish a generalisable and flexible mapping of known cyber and information security risks present during third-party model evaluations. The paper is the outcome of the SAFA-TF,[1] an initiative that gathered evaluators, AI labs, cybersecurity experts and AI governance researchers to map risks linked to third-party access and match them with appropriate countermeasures. In this paper, the authors present an Access–Risk Matrix outlining the existing and potential threats to access, proposed mitigations for these risks, and the respective stakeholder group responsibilities associated with each. Underpinning this effort is the objective to inform a debate on developing a risk-informed framework for evaluator access, which enables meaningful third-party access and manages risks proportionately.

1. RUSI, 'Secure Access to Frontier AI Taskforce', 2026, <https://www.rusi.org/explore-our-research/projects/secure-access-frontier-ai-taskforce>, accessed 21 April 2026.

The Multistakeholder Landscape of Frontier AI Evaluations

As frontier AI models expand in their capability and application, it is key that their evolution remains grounded in safety and security safeguards to prevent them from being misused or repurposed to conduct cyberattacks, terrorist attacks and other harmful activities. In the past, companies have been increasingly publishing their Frontier AI Safety Frameworks[2] and threat intelligence reports to share[3] their approaches[4] to managing risks and potential threats.[5] More importantly, AI labs such as Anthropic, Google DeepMind, OpenAI, Meta and others have established internal safety and security teams to conduct such tests,[6] but have also increasingly relied on external evaluators to assess, among other things, the capabilities (including dangerous ones) of frontier models prior to or following their deployment. While internal evaluations are critical, the pace, scale and commercial pressures of frontier AI development mean that, despite good intentions, developers might be susceptible to blind spots.

A growing ecosystem of independent evaluation organisations has emerged alongside AI safety and security benchmarking from AI labs. These third-party evaluations support labs with safety and security assessments while providing additional assurance over the performance of the frontier model. These external evaluators constitute a nascent field of organisations that include, but are not restricted to:

- Apollo Research, which specialises in assessments of scheming behaviours (when advanced frontier AI systems covertly pursue misaligned objectives); [7]
- METR, which focuses on different capability assessments including evaluating how much an AI can carry out substantial general-purpose tasks;[8]
- The AI Verification & Evaluation Research Institute (AVERI), which looks at auditing for security and safety; [9]

2. Yoshua Bengio et al., 'International AI Safety Report 2026', 3 February 2026, <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>, accessed 15 February 2026.
3. Google, 'Secure AI Framework (SAIF)', <https://safety.google/intl/en_ca/safety/saif/>, accessed 10 January 2026.
4. Open AI, 'Open AI, Strengthening Cyber Resilience as AI Capabilities Advance', 10 December 2025, <https://openai.com/index/strengthening-cyber-resilience>, accessed 16 April 2026.
5. Nicholas Carlini et al., 'Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days', *Red Anthropic,* 5 February 2026, <https://red.anthropic.com/2026/zero-days/>, accessed 6 February 2026.
6. Google DeepMind, 'Evals: Explore our Comprehensive Evaluations Across AI Capabilities', <https://deepmind.google/research/evals/>, accessed 2 February 2026.
7. Bronson Schoen et al., 'Stress Testing Deliberative Alignment for Anti-Scheming Training', Apollo Research, 19 September 2025, <https://www.apolloresearch.ai/research/stress-testing-deliberative-alignment-for-anti-scheming-training/>, accessed 20 January 2026.
8. METR, 'About METR', <https://metr.org/about>, accessed 21 January 2026.
9. AI Verification & Evaluation Research Institute (AVERI), 'About', <https://www.averi.org/about>, accessed 21 January 2026.

- Irregular, which concentrates in cybersecurity evaluations, among others.[10]

Industry coordination through the Frontier Model Forum[11] has contributed to further codifying the role of external evaluators, identifying three core functions for third-party assessments, which complement rather than replace, internal processes:

- confirming the soundness of developer-led evaluations;
- applying independent methods to test the robustness and safety claims of developer-conducted evaluations;
- supplementing internal capabilities and capacity of developers to perform internal assessments.

Governments have also responded by creating AI safety and security institutes and a network of such organisations with the aim of facilitating AI safety and security evaluations.[12] The UK's AI Security Institute (AISI) assessed 30 models over a period of two years (2023–25) across several security-critical domains.[13] The US Center for AI Standards and Innovation conducted pre-deployment evaluations and established voluntary testing agreements with frontier developers. In 2025, their mandate was refocused on national security, whereby they would, among other things, lead evaluations and assessments 'of potential security vulnerabilities and malign foreign influence arising from use of adversaries' AI systems, including the possibility of backdoors and other covert, malicious behavior'.[14]

Regulatory frameworks and governmental policy papers have only begun to attempt to codify these expectations, to varying degrees of detail and success. The Bletchley Summit declaration in 2023 acknowledged that evaluations and assessments of the capabilities and risks of AI systems (such as model evaluations and/or red teaming) should be key components of AI governance regimes.[15] But beyond aspiration and the AISI network, such national or third-party metrics and methodologies for evaluations are socialised but not always shared among stakeholders. Even so, public and policymaker knowledge on how frontier models are evaluated, developed, deployed

10. Irregular, 'The Next Generation of Cyber Evaluations', 19 November 2025, <https://www.irregular.com/publications/next-generation-of-cyber-evals>, accessed 21 January 2026.
11. Frontier Model Forum, 'Technical Report: Third-Party Assessments', 4 August 2025, <https://www.frontiermodelforum.org/technical-reports/third-party-assessments/>, accessed 21 January 2026.
12. EU, 'First Meeting of the International Network of AI Safety Institutes', 20 November 2024, <https://digital-strategy.ec.europa.eu/en/news/first-meeting-international-network-ai-safety-institutes>, accessed 21 January 2026.
13. AI Security Institute (AISI), 'Frontier AI Trends Report', December 2025, <https://www.aisi.gov.uk/frontier-ai-trends-report/pdf>, accessed 22 January 2026.
14. NIST, 'Center for AI Standards and Innovation (CAISI)', <https://www.nist.gov/caisi>, accessed 22 January 2026.
15. Anka Reuel et al., 'Open Problems in Technical AI Governance', *Transactions on Machine Learning Research* (April 2025),<https://arxiv.org/abs/2407.14981>, accessed 22 January 2026.

and safeguarded remains limited.[16] The passing of the EU AI Act in 2024[17] introduced evaluation obligations at two distinct layers:

- At the system level, AI systems deployed in high-risk use cases are subject to conformity assessments by providers or designated notified bodies.
- At the model level, where providers of general-purpose AI models classified as posing systemic risk face obligations, including dangerous capability evaluations, adversarial testing and systemic risk assessments.

The General-Purpose AI Code of Practice (CoP) operationalises these model-level requirements, committing signatories to granting independent external evaluators 'sufficient access, information, time and resources' to their models, including access to internal model components and unmitigated versions where appropriate.[18] It is this model-level evaluation regime – albeit not exclusive to the European context – that is the primary focus of this paper, along with the access, security and methodological challenges it raises.

The Challenge: Managing Access and Security Risks

Enabling meaningful third-party evaluation might require granting external parties access to some of the most sensitive information about these frontier models. A core concern for AI companies is guarding against IP leakage, reverse engineering and model compromise. Irregular or malicious access to proprietary model elements, such as architecture, training data, model weights or inference infrastructure, could compromise substantial research and development investments. The potential for exploitation during third-party access, whether by financially motivated or state-sponsored actors, represents a material risk. Additionally, the potential for unauthorised manipulation presents both reputational and operational threats, including model and data poisoning attacks that could degrade model performance and safety. The lack of consensus on best practices or standards equally contributes to these concerns – perceived cyber and information security risks may play a significant role in stifling the third-party evaluation ecosystem.

The rapid proliferation of evaluation types in recent years – a natural response to the expanding application domains of frontier models – has not been matched by corresponding clarity on access requirements or their associated security risks.

16. Bengio, 'International AI Safety Report 2026'.
17. EU, 'Regulation (EU) 2024/1689 (EU AI Act)', June 2024, <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>, accessed 22 January 2026.
18. European Commission, 'The General-Purpose AI Code of Practice', 10 July 2025, <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>, accessed 21 January 2026.

Despite significant developments in AI safety and security research, no authoritative or widely accepted framework maps evaluation types to access modalities and their security implications. The CoP, for example, calls for 'appropriate access' but does not define what this means in practice.[19] No common framework currently exists for describing different types and levels of evaluator access, and research has suggested that the interpretation of 'sufficient access' remains ambiguous.[20] While this paper does not seek to define 'sufficient access', it is critical that policy discussions have an evidence-driven and concrete understanding of risks linked to certain forms of access.

The absence of a shared framework has practical consequences for both evaluators and developers and, most importantly, highlights the need for a collective effort to map risks linked to access with the aim of devising shared mitigation strategies. The UK government's 'Roadmap to Trusted Third-Party AI Assurance' paper, published in September 2025, notes that evaluators 'struggle to access the information they need to conduct effective assurance of AI systems' due to a range of challenges: concerns about sharing commercially sensitive information, security and privacy risks deriving from access; and lack of established practices from AI labs to make information accessible to evaluators.[21] The EU's AI Act acknowledges that the third-party evaluation ecosystem is still maturing, recognising the need to build capacity and expertise as the assurance sector scales to meet the demands of evaluating cutting-edge AI models.[22] Researchers have demonstrated significant variance among major AI companies in access policies, with some embracing transparency while others maintain near-zero third-party access to non-public models.[23] This fragmentation has practical consequences, one of which is that insufficient or inadequate release of information could increase risk, while overly stringent policies could 'centralise power and stifle critical research'.[24] For developers, the heightened risk sensitivity regarding frontier AI models has driven conservative approaches to third-party access. Stringent access restrictions have constrained where and how independent researchers are empowered to conduct meaningful tests.[25]

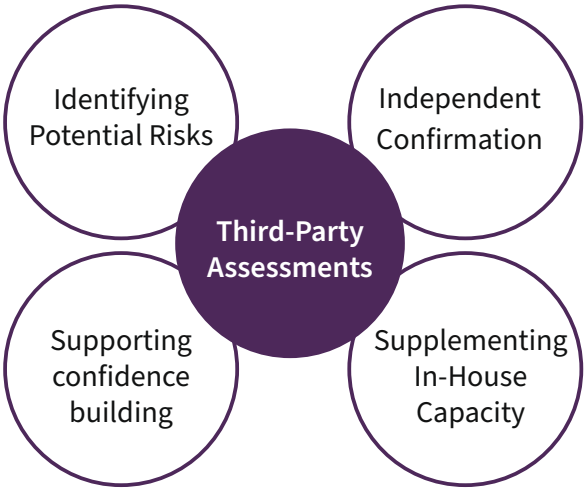
19. Jacob Charnock et al., 'Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations', arXiv preprint arXiv:2601.11916, 21 January 2026, <https://arxiv.org/abs/2601.11916>, accessed 22 January 2026.
20. Miles Brundage et al., 'Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies', arXiv preprint arXiv:2601.11699, 16 January 2026, <https://arxiv.org/abs/2601.11699>, accessed 5 February 2026.
21. Department for Science, Innovation and Technology, 'Trusted Third-Party AI Assurance Roadmap', policy paper, 3 September 2025, <https://www.gov.uk/government/publications/trusted-third-party-ai-assurance-roadmap/trusted-third-party-ai-assurance-roadmap>, accessed 5 February 2026.
22. EU AI Act, 12 July 2024, <https://artificialintelligenceact.eu/>, accessed 6 May 2026.
23. Shayne Longpre et al., 'Position: A Safe Harbor for AI Evaluation and Red Teaming', in *Proceedings of the 41st International Conference on Machine Learning* (ICML'24) (PMLR, 2024, Vol. 235).
24. Edward Kembery, Ben Bucknall and Morgan Simpson, 'Position Paper: Model Access Should be a Key Concern in AI Governance', arXiv preprint arXiv:2412.00836, 1 December 2024, <https://arxiv.org/abs/2412.00836>, accessed 5 February 2026.
25. Kevin Klyman et al., 'Safeguarding Third-Party AI Research', Stanford University Human-Centered Artificial Intelligence, February 2025, <https://hai.stanford.edu/assets/files/hai-policy-brief-safeguarding-third-party-ai-research.pdf>, accessed 5 February 2026.

Strengthening Assessments: Why Third-Party Access Matters

What Do Third-Party Evaluators Do?

Third-party assessments support greater AI transparency and accountability in at least four ways: providing independent confirmation; supplementing in-house capacity to conduct such evaluations; supporting confidence-building of the secure and safe deployment of frontier models; and identifying potential risks through a constantly evolving set of methodologies which includes, for example, the Frontier Model Forum's work on the topic (Figure 1).

Figure 1: Third-Party Evaluation Functions



Sources: The authors. Diagram reflective of previous research, such as the ones seen in the following sources: Frontier Model Forum, 'Third-Party Assessments', 4 August 2025, <https://www.frontiermodelforum.org/technical-reports/third-party-assessments/>, accessed 21 January 2026; Anthropic, 'Third-Party Testing as a Key Ingredient of AI Policy', 25 March 2024, <https://www.anthropic.com/news/third-party-testing>, accessed 13 April 2026.

Research has also suggested that such evaluations can additionally enhance AI labs' compliance with their own safety frameworks, as well as external ones, including binding and non-binding frameworks.[26] They can assure external stakeholders through independent evaluation and validation, and it can also lessen the internal stakeholder burden by reassuring them that they are adequately adhering to the safety framework.

Third-party evaluations are a complementary element to existing internal developer-led evaluations. Frontier model developers perform extensive internal testing with deeper system visibility and operational context. Third-party evaluators play an important role in supporting the innovation of frontier AI models by challenging, validating and providing diverse perspectives on their security and safety.

The Importance of Access

Third-party evaluators need a certain level of access to the frontier AI model to properly assess its performance, capabilities, systemic risks and the security and safety functions. As previously mentioned, this access enables evaluators to analyse the model's performance in several ways: whether the model is performing as expected, in a robust manner, and with the purpose of evaluating claims about the model's security and safety.

However, there is a need to decide how, when and under what conditions a model should or could be evaluated by a third party with the aim of enhancing the security and safety of its operation and deployment. Based on previous research, the access required to conduct third-party evaluations of models tends to vary across research areas and uses.[27] Benjamin S Bucknall and Robert F Trager's taxonomy of model access identifies multiple dimensions – including sampling, inspection, fine-tuning and modification – each carrying different security implications.[28] Yet, as they note, there remains 'little clarity regarding the functionality that would be most useful for researchers if provided through such a service'.

26. Aidan Homewood et al., 'Third-Party Compliance Reviews for Frontier AI Safety Frameworks', arXiv preprint arXiv:2505.01643, 3 May 2025, <https://arxiv.org/abs/2505.01643>, accessed 13 April 2026.
27. Stephen Casper et al., 'Black-Box Access is Insufficient for Rigorous AI Audits', in the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), New York, 3–6 June, 2024, <https://arxiv.org/abs/2401.14446>, accessed 2 February 2026.
28. Benjamin S Bucknall and Robert F Trager, 'Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements', Oxford AI Governance Initiative, October 2023, <https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf>, accessed 12 January 2026.

Research has consistently reflected that sufficient access to models and underlying, internal information is critical. Edward Kembery, Ben Bucknall and Morgan Simpson argue that there is a necessary balance between granting sufficient model access and accepting and mitigating the risks of access mismanagement.[29] Below are some examples of the types of assessments conducted by evaluators: [30]

- **Dangerous capability assessment**: Understanding if a model can enable large-scale harm before safety measures are implemented (for example, supporting chemical, biological, radiological and nuclear defence weapons development, automating cyberattacks[31] and operations and/or conducting other dangerous autonomous operations).
- **Propensity assessment**: Testing whether frontier models might be pursuing misaligned goals,[32] exhibiting unexpected and/or undesirable covert behaviours[33] such as deception, sycophancy, reward hacking or sabotage, among others.[34]
- **Safeguard testing**: Determining whether safety measures and mitigations hold under adversarial pressure and attempts to circumvent them (for example, jailbreaks or prompt injection attacks).
- **Human uplift evaluations**: Assessing how frontier models can be used by malicious actors to conduct real-life, harmful tasks.[35]

The levels or degrees of access to frontier AI models, however, remain subject to debate and non-standardised. As identified by Miles Brundage and others, there is no consensus interpretation of 'minimum access' regarding third-party evaluations.[36] This significantly complicates model access governance, decreasing the scope of access and inhibiting comprehensive third-party evaluation.[37] Experts note that the lack of adequate access to AI systems represents a significant vulnerability to the AI assurance ecosystem.[38]

29. Kembery, Bucknall and Simpson, 'Position Paper'.
30. Department for Science, Innovation and Technology (DSIT), 'AI Safety Institute Approach to Evaluations', 9 February 2024, <https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations>, accessed 2 February 2026.
31. Irregular, 'Evaluating GPT-5.2 Thinking: Cryptographic Challenge Case Study', 11 December 2025, <https://www.irregular.com/publications/spell-bound-technical-case-study>, accessed 20 February 2026.
32. Schoen et al., 'Stress Testing Deliberative Alignment for Anti-Scheming Training'.
33. Alexander Meinke et al., 'Frontier Models are Capable of In-context Scheming', arXiv preprint arXiv:2412.04984, 14 January 2025, <https://arxiv.org/abs/2412.04984>, accessed 3 March 2026.
34. OpenAI, 'Detecting and Reducing Scheming in AI Models', 17 September 2025, <https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/>, accessed 3 March 2026.
35. AI Safety Institute, 'AI Safety Institute Approach to Evaluations', 9 February 2024, <https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations>, accessed 3 March 2026.
36. Brundage et al., 'Frontier AI Auditing', arXiv preprint arXiv:2601.11699, 16 January 2026, <https://arxiv.org/abs/2601.11699>, accessed 5 February 2026>.
37. Kembery, Bucknall and Simpson, 'Position Paper: Model Access should be a Key Concern in AI Governance', arXiv preprint arXiv:2412.00836, 1 December 2024, <https://arxiv.org/abs/2412.00836>, accessed 13 April 2026.
38. Inioluwa Deborah Raji et al., 'Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance', presented at the 5th Annual ACM/AAAI AI Ethics and Society (AIES) Conference, August 2022, <https://arxiv.org/abs/2206.04737>, accessed 13 April 2026.

Table 1 provides an indicative list of access types and data points associated with them. While access is thereby depicted in terms of the broader categories of data which evaluators are accessing (for example, query and inference, model internals, environment and scaffolding), each of these varies in the depth and sensitivity of access. Such a degree of access can range from black-box querying – which only provides the evaluator with the capacity to query the model and observe outputs without any access to information about the model's internal workings – to full white-box access. This paper draws on a taxonomy of access levels developed by experts in this field and reflect on their direct link to specific data points typically required for certain types of access.[39]

- **AL1 (black-box access)**: Black-box access with minimal information. The evaluator can query the model and observe outputs but has no visibility into internal workings – no access to architecture, weights, activations or training data – and receives only basic documentation. Evaluators typically have access only to query and inference outputs and basic configuration information. This level of access enables basic external assurance and vulnerability detection.
- **AL2 (grey-box access)**: Grey-box access with substantial information. The evaluator receives partial access to internal model components, such as log-probabilities of output tokens, the ability to fine-tune on custom datasets or modify sampling parameters and more detailed documentation about the training methodology, safety mitigations and evaluation results. This level of access can help improve capability elicitation and vulnerability identification.
- **AL3 (white-box access)**: White-box access with comprehensive information. The evaluator has full access to the model's internal components (in other words, weights, architecture, activations, gradients and training data) and comprehensive development documentation, such as design rationale, internal evaluation findings and security architecture. This level of access probably provides the highest level of accuracy in risk assessments, given the level of visibility and documentation accessed by evaluators.

For example, a black-box level of access enables a degree[40] of testing but features critical shortfalls,[41] such as implicit bias in how evaluators' inputs are crafted and a limited number of inputs that can be tested. Furthermore, Stephen Casper and others argue that explanations for black-box analyses may misidentify causal relationships between model inputs and outputs, largely due to not having crucial context regarding model design.[42] To better identify relationships and hedge against input-bias result

39. Jacob Charnock et al., 'Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations', arXiv preprint arXiv:2601.11916, 17 January 2026, <https://arxiv.org/abs/2601.11916>, accessed 13 April 2026.
40. Miles Brundage et al.,"Frontier AI Auditing', arXiv preprint arXiv:2601.11699, 16 January 2026, <https://arxiv.org/abs/2601.11699>, accessed 5 February 2026.
41. Casper et al., 'Black-Box Access is Insufficient for Rigorous AI Audits'.
42. *Ibid*.

skew, the authors argue that white-box and out-of-the-box access may be required to conduct rigorous and trustworthy external assessments. This expanded access includes models' white box, inclusive of more discrete features such as model architecture, decision tree logic, training processes and safety and control filters. Each progression in access level enables more rigorous evaluation but carries greater security risk – a trade-off that the Access–Risk Matrix in the fourth chapter addresses directly.

Table 1: Types of Access Required to Conduct Evaluations

Access Type	Description	Required/Relevant Data Points	Access Level (AL)
Query and Inference	Access to model outputs generated in response to prompts, including processed and raw outputs	Final output; raw output; logits; Chain of Thought (summary and raw); and refusal patterns	AL1 (final output, refusal patterns) AL2 (logits, raw Chain of Thought) AL3 (raw activations and outputs)
Model Internals (Read)	Read-only access to the internal computational structures and learned representations within the model	Activations (read); weights (read); and architecture states	AL2 (partial access of selected activations, limited architecture information) AL3 (full access to model weights, architecture and internal states)
Model Internals (Write)	Access to modify (write) model's internal parameters, representations and/or safety mechanisms	Activation modification; weights (write/fine-tuning); safety control modification	AL2 (fine-tuning) AL3 (full write access)
Configuration	Access to system-level settings that shape model behaviour without modifying weights	System prompts; environment parameters	AL1 (basic documentation, if disclosed) AL2 (system prompts, environment parameters) AL3 (full-configuration access including deployment settings)
Training and Evaluation Data	Access to the datasets, results and methodologies used to train and assess the model	Training metadata; pre-training samples; post-training data; evaluation data and results; threat models	AL1 (minimal documentation) AL2 (metadata, evaluation results) AL3 (full training data, pre-/post-training samples and threat models)
Environment and Scaffolding	Access to the infrastructure, external connections and tools that extend model capabilities	Computational tools (code execution capabilities); external system access (API, network access); API credentials (authentication tokens); model name; input/output filters	AL1 (model name, basic API access and input/output filters) AL2 (code execution capabilities, extended API access and authentication tokens) AL3 (full infrastructure access and network configurations)
Compute/ Inference Infrastructure	Access to the hardware, computational resources and additional infrastructure used to train, fine-tune and further develop the model	Inference server architecture; hardware specifications; network specifications; power and cooling specifications	AL2 (inference server configurations) AL3 (full infrastructure access, hardware specifications and network specifications)

Source: The authors.

Associated Threats

This chapter identifies seven high-level security risk categories and how these risks connect to specific data and system access points.

Considering how nascent the field of AI safety and security is, and specifically the role and access of third-party evaluators to these models, there are a few elements to consider in this chapter.

The ecosystem for structured third-party evaluation is still developing. This means that formalised relationships, shared understandings of access, methodologies and types of assessments have only recently begun to scale (and are constantly evolving). This poses a challenge to systematically mapping risks in a structured manner and offers an opportunity for further research. Disclosure of incidents linked to frontier models remains limited. Despite extensive cybersecurity and AI safety research, there is still no comprehensive framework for public reporting of incidents involving the misuse and exploitation of models – whether identified through real-world use, internal testing, adversarial red teaming, third-party evaluations or academic research. However, not all AI misuse and security risks are strictly cybersecurity risks – they encompass and intersect with broader information security and safety risks.

Moreover, some of the threats are hypothetical. Some of those outlined below represent theoretically viable exploitation pathways based on the underlying design of current AI model architectures. This makes it difficult to determine which security risks are still theoretical and which have been substantiated through research, testing or documented use. With that in mind, the security risk categories below draw on both academic research and public reports available at the time of writing. Moreover, the information asymmetry between developers, evaluators and researchers means that empirical evidence for some risks is still not entirely publicly available. This, however, does not diminish their relevance; rather, it underscores the importance of and need for mapping such risks.

Explanation: Adversarial Exploitation *versus* Unintentional Compromise versus Model Misbehaviour

Security risks associated with third-party access to frontier AI models can be understood through at least three categories, which merit further clarification. While this paper primarily focuses on the first two – and only on human agency in enabling or amplifying those risks – all three categories interact and can result in an aggregated risk.

Adversarial exploitation: Deliberate actions by human actors such as state-sponsored groups, criminal groups, commercial competitors, or insider threats who intentionally exploit access to frontier models with the purpose of, for example, stealing, manipulating or weaponising them.

Unintentional compromise: Security or operational failures that occur without malicious intent but that can result in further exploitation by humans or AI systems. No adversary is involved, but the consequences can still be severe regardless of whether the outcomes were intentional or unintentional.

Model misbehaviour: When the model behaves in a way that is misaligned with what had been intended by the developer. Misbehaviour, such as scheming (covertly pursuing misaligned goals), is not caused by an external actor but emerges from the model's training or architecture.

Security Risk Categories

Below are outlined key security risk categories drawn from a series of RUSI SAFA-TF meetings. These risks are separated into four clusters: structural exploitation, adversarial probing, operational failures and outcomes. While non-exhaustive and not mutually exclusive, they represent some of the community's main security concerns. The risk categories identified are high level, each encompassing several specific means of access and threat vectors. The examples provided for each category are not necessarily specific to third-party evaluations; rather, they are recent, real-world examples of each risk. These categories reflect security risks that might derive either from intentional or unintentional action (in other words, accidental exposure). The next section shows how these different security risk categories manifest (or might do so) in the context of third-party access to frontier models.

Structural Exploitation

Model Theft: An adversary acquires a functional copy of the model, either through direct exfiltration of weights or other core artefacts or distillation attacks.

- **Primary Threat Actors**: State-sponsored or competitors.
- **Examples**: In 2025, OpenAI stated that the China-based AI company DeepSeek may have used expansive automated querying to distil OpenAI models to build their own models.[43] The allegations do not include illegal or improper access but rather abuse of OpenAI's API. The US government indicated that they were probing the use of distillation by Chinese firms in an attempt to prevent the development of 'copycat models'.[44] Such distillation-based theft is one of dozens of attack vectors documented against model weights.[45]

Capability Reconstruction: An adversary obtains sufficient process knowledge (training methodology, data pipelines and alignment techniques) to replicate a model or develop a similar one.

- **Primary Threat Actors**: State-sponsored, criminal groups/actors, competitors or insiders.
- **Examples**: In 2025, a former Google employee was indicted by the US Department of Justice for economic espionage and theft of trade secrets.[46] The former employee allegedly stole information about the 'hardware infrastructure and software platform' underpinning Google's AI model-training supercomputing data centres.[47] While details about the case are insufficient to determine whether theft has led to full model replication, third-party safety researchers and adversarial machine learning scientists such as Nicholas Carlini and others[48] have demonstrated the ability to at least partially reconstruct training data, alignment and safety behaviours of GPT-2, as well as the capacity to steal part of a production language model.[49]

43. Cade Metz, 'OpenAI Says DeepSeek May Have Improperly Harvested Its Data', *New York Times*, 29 January 2025.
44. Andrea Shalal, David Shepardson and Kanishka Singh, 'White House Evaluates Effect of China AI App DeepSeek on National Security', *Reuters*, 28 January 2025.
45. Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (Santa Monica, CA: RAND Corporation, 2024).
46. US Department of Justice, 'Superseding Indictment Charges Chinese National in Relation to Alleged Plan to Steal Proprietary AI Technology', press release, 4 February 2025, <https://www.justice.gov/usao-ndca/pr/superseding-indictment-charges-chinese-national-relation-alleged-plan-steal>, accessed 13 April 2026.
47. Jonathan Stempel, 'Ex-Google Engineer Faces New US Charges He Stole AI Secrets for Chinese Companies', *Reuters*, 5 February 2025.
48. Nicholas Carlini et al., 'Extracting Training Data from Large Language Models', arXiv preprint arXiv:2012.07805, December 2020, <https://arxiv.org/abs/2012.07805>, accessed 13 April 2026.
49. Nicholas Carlini et al., 'Stealing Part of a Production Language Model', in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, Vienna, 21–27 July 2024), pp. 5680–705, <https://arxiv.org/abs/2403.06634>, accessed 13 April 2026.

Model Manipulation: When an adversary exploits access through data poisoning, backdoor insertion, adversarial fine-tuning or direct modification of safety controls to alter the model, it behaves differently (or pursues misaligned objectives) from what had been intended by the developer. Such actions might therefore result in degraded safety performance, covert harmful functionality or compromised integrity.

- **Primary Threat Actors**: State-sponsored, criminal groups/actors, insiders or competitors.
- **Examples**: Launched in 2023, WormGPT is a black-hat GPT tool used by criminals and other malicious actors to support their tasks.[50] The model was built on the open source language model GPT-J 6B. The creator of WormGPT claimed to 'have fine-tuned this accessible foundation model using specialized, confidential and malicious datasets with a specific emphasis on malware-related data'.[51] This is an example of how the resulting tool had purposefully been developed without the adequate safeguards of other AI models and therefore an 'uncensored' alternative to mainstream LLMs designed for phishing, malware generation and social engineering.

Adversarial Probing

Jailbreak/Safety Bypass: An adversary exploits vulnerabilities to bypass safety controls and access model capabilities through prompt injection and other methods/tools. This category encompasses both direct attacks on safety controls (jailbreaks) and the exploitation of model behaviour through injected instructions (prompt injection). While jailbreaks target the model's safety training via query access, prompt injection exploits the model's interaction with external content – a risk that scales with the model's access to tools, documents and external systems.

- **Primary Threat Actors**: State-sponsored or criminal groups/actors.
- **Examples**: In early 2025, independent third-party researchers from Aim Security Labs discovered a vulnerability in Microsoft 365's Copilot AI, dubbed 'EchoLeak'.[52] An external email containing 'implants hidden instructions' was processed by Copilot, leading it to respond by leaking private data, allowing malicious actors to exfiltrate the data onto an external server.[53] The vulnerability enabled attackers to access 'internal files or messages' through the exploited Microsoft Copilot AI. While the EchoLeak case is not a jailbreak per se – as the former would mean a user deliberately crafting inputs to bypass

50. Mohamed Fazil Mohamed Firdhous et al., 'WormGPT: A Large Language Model Chatbot for Criminals', in 24th International Arab Conference on Information Technology (ACIT 2023), Ajman, Jordan, 6–8 December 2023, <https://ieeexplore.ieee.org/document/10453752>, accessed 13 April 2026.
51. Unit 42, 'The Dual-Use Dilemma of AI: Malicious LLMs', 25 November 2025, <https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/>, accessed 13 April 2026.
52. Itay Ravia, 'Breaking Down "EchoLeak", the First Zero-Click AI Vulnerability Enabling Data Exfiltration from Microsoft 365 Copilot', *Cato Networks* blog, 31 May 2025, <https://www.catonetworks.com/blog/breaking-down-echoleak/>, accessed 13 April 2026.
53. Pavan Reddy and Aditya Sanjay Gujral, 'EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System', arXiv preprint arXiv:2509.10540, 6 September 2025, <https://arxiv.org/abs/2509.10540>, accessed 13 April 2026.

security – it is still a safety bypass example via 'indirect' (email with hidden instructions) prompt injection.

Explanation: Distinguishing Between Security Risks

While many adversarial tactics are used to misalign model behaviour, it is important to understand how they do so differently and which elements of the frontier model they seek to modify and/or hijack.

Jailbreak: The user crafts inputs that trick the model into ignoring its safety training. The model itself is unchanged, and the effect resets after the session.

Prompt injection: An attacker embeds hidden instructions in the content the model processes (for example, emails and documents), causing it to follow those instructions instead of doing what it was asked to do. Similarly to jailbreaking, the model itself is unchanged.

Model manipulation: An adversary with deeper access alters the model itself through data poisoning, backdoor insertion or modification of safety controls. Unlike jailbreaking or prompt injection, the change is persistent: the model is structurally different, and all subsequent users are affected. This is distinct from jailbreaking or prompt injection, which focus on the exploitation of the model's behaviour rather than, in this case, the model itself.

Credential Compromise [cross-cutting vector]: An adversary exploits third-party access after having obtained their credentials (through phishing, break or insider compromise, allowing them to escalate privileges).

- **Primary Threat Actors**: State-sponsored or criminal groups/actors.
- **Examples**: In August 2025, Google's Threat Intelligence Group reported threat actors had successfully stolen 'OAuth' and refresh tokens used by Salesloft's Drift AI Chatbot, a trusted third-party application integrated into Salesforce customer-specific environments.[54] These tokens served as credentials by which the threat actors were able to gain unauthorised access to private company data via the AI chatbot, and exfiltrate 'large volumes' of Salesforce customer data. According to Salesloft, the attackers' objective was credentials, 'focusing on sensitive information like AWS access keys, passwords and Snowflake-related access tokens'. [55]

54. Austin Larsen et al., 'Widespread Data Theft Targets Salesforce Instances via Salesloft Drift', Google Cloud blog, 26 August 2025, <https://cloud.google.com/blog/topics/threat-intelligence/data-theft-salesforce-instances-via-salesloft-drift>, accessed 13 April 2026.
55. Salesloft, 'Drift/Salesforce Security Update', 20 August 2025.

Operational Failures

Accidental Exposure: Sensitive model or model-related information that is unintentionally/accidentally exposed due to an error (human or technical) or inadequate protection protocols, processes and/or procedures.

- **Primary Threat Actors**: Not applicable as it is non-adversarial and non-intentional.
- **Examples**: In February 2023, Meta distributed LLaMA, a frontier AI model under development, to approved external researchers to conduct safety and security safeguards. However, LLaMa was distributed without 'restricting access to the underlying data, software, and model'.[56] The full model, inclusive of weights, was published publicly on BitTorrent. While META did not intend to publicise its model information, a combination of human and technical errors, compounded by inadequate protections, led to the exposure of sensitive materials.

Access Persistence [cross-cutting vector]: Third-party access remains active after the authorised period is over and is exploited by adversaries who maintain unauthorised use.

- **Primary Threat Actors**: State-sponsored, criminal groups/actors and compromised third-party evaluators.
- **Examples**: In July 2024, *Business Insider* reported that a company with a third-party contract with Anthropic had left an open-access folder containing guidance on which websites to mine in the process of 'fine-tuning' Anthropic's AI.[57] While this was a misconfiguration, threat actors could exploit such a vulnerability (while undetected) to ensure continued access to the folder. According to *Business Insider*'s report, there is no evidence of malicious use of the contents of the leaked folder.

Outcomes

Weaponisation: Adversary applies model capabilities – obtained through legitimate or illegitimate access – to conduct real-world harm, such as automating cyber-reconnaissance or scaling disinformation.

- **Primary Threat Actors**: State-sponsored or criminal groups/actors.
- **Examples**: Anthropic disclosed that a cybercriminal used Claude code to develop, market and distribute multiple ransomware variants with advanced evasion capabilities.[58] The threat actor leveraged Claude's code execution environment to 'automate reconnaissance, credential harvesting, and network penetration at scale,

56. Richard Blumenthal and Josh Hawley, Letter to Mark Zuckerberg regarding Meta's LLaMA model, 6 June 2023, <https://www.blumenthal.senate.gov/imo/media/doc/06062023metallamamodelleakletter.pdf>, accessed 13 April 2026.
57. Charles Rollet, 'Here's the List of Websites Gig Workers Used to Fine-Tune Anthropic's AI Models. Its Contractor Left It Wide Open', *Business Insider*, 23 July 2025, <https://www.businessinsider.com/anthropic-surge-ai-leaked-list-sites-2025-7>, accessed 13 April 2026.
58. Anthropic, 'Detecting and Countering Misuse of AI: August 2025', 27 August 2025, <https://www.anthropic.com/news/detecting-countering-misuse-aug-2025>, accessed 13 April 2026.

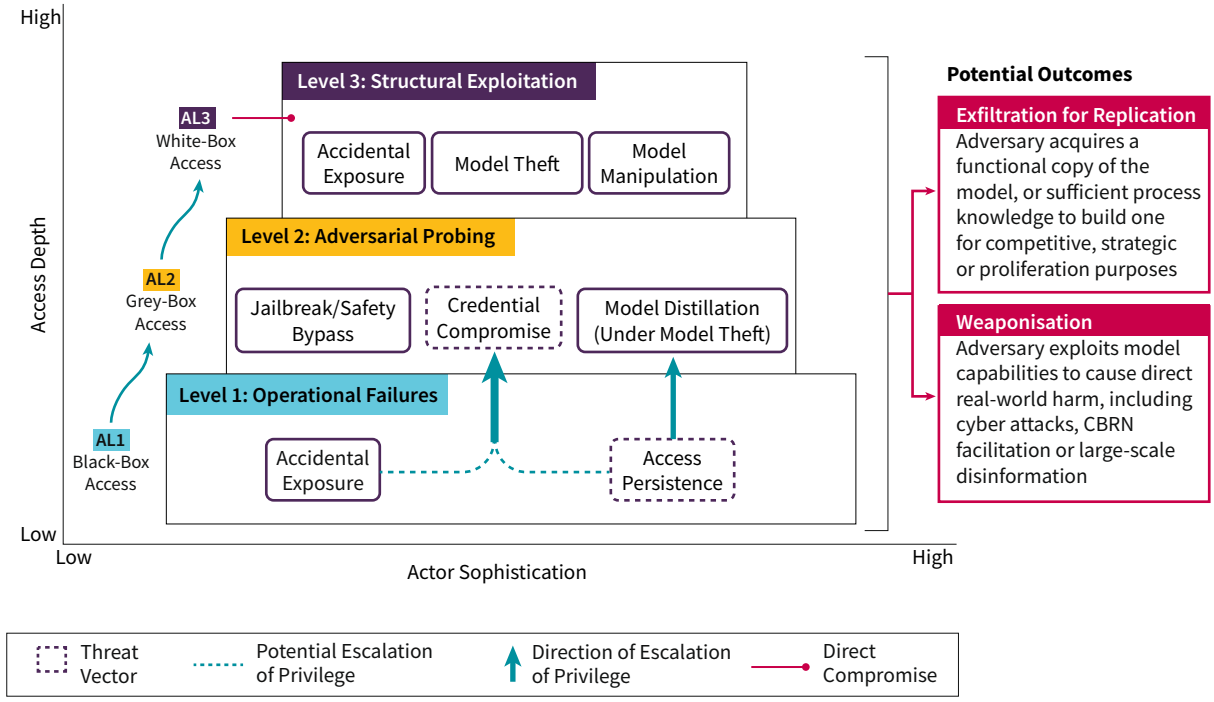
potentially affecting at least 17 distinct organizations' across 'government, healthcare, emergency services, and religious institutions'.[59]

Operationalising the Security Risks Taxonomy for Third-Party Access

This section connects the high-level security risks taxonomy with potential levels of third-party access to frontier models in two ways. First, it provides a visual representation of direct and indirect pathways of access exploitation that may lead to malicious security outcomes such as full model exfiltration and other forms of weaponisation. Second, it provides a subsequent representation of the connection between the access types outlined in Table 1 and specific data points and security risks, giving an incrementally more granular understanding of the link between access and risks.

Exploitation Pathways

Figure 2: Pathways of Exploitation of Access to Frontier Models



Source: The authors. Graphic: Alex Whitworth.

59. Anthropic, 'Threat Intelligence Report: August 2025', August 2025, <https://www.anthropic.com/news/detecting-countering-misuse-aug-2025>, accessed 13 April 2026.

Drawing on the degrees of access outlined in Table 1 (AL1-3), Figure 2 represents how security risks such as those linked to structural exploitation, adversarial probing and operational failures can be exploited either directly or through lateral movement and privilege escalation.

The illustration combines the security risks taxonomy with the levels of access, reflecting on how potential escalation pathways and direct compromise of access types could lead to malicious outcomes.

Connecting Access Types with Security Risks

Experts on the SAFA-TF flagged that a significant number of data points, many of which are quickly evolving, remain underexplored in relation to the threats they may enable. Table 2 demonstrates the association between a non-exhaustive and summarised list of certain data points and the threats they could enable through access, as identified by the SAFA-TF.

This mapping is illustrative rather than exhaustive. Comprehensive threat modelling would need to be tailored to specific evaluation types and access levels. Table 2, along with Figure 1, is only one step towards going beyond broad categories of access and understanding that malicious actors can exploit certain types of data to advance through what is known as the 'Cyber Kill Chain' (reconnaissance, weaponisation, delivery, exploitation, installation, command and control, actions and objectives) within a closed frontier AI ecosystem.

Three observations follow from this analysis. First, the threat landscape facing third-party access is considerably more nuanced – and more of a gradient – than a binary or ideal-type of low-risk black-box access and high-risk white-box access. As this section shows, access can enable distinct and, in some cases, compounding attack pathways. Second, cross-cutting vectors such as credential compromise and access persistence are not unique to AI security but are structural concerns in cybersecurity and information security, and, as in other areas, they amplify risks across the board. Third, access types that might be considered low risk can still serve as enablers of more serious exploitation. These findings reflect that third-party access requires not a blanket restriction but a shared, risk-informed framework capable of calibrating protections to the specific threats that different evaluation contexts present.

Table 2: Access Types, Data Points and Associated Security Risks

Access Type	Data Points	Associated Security Risks
Query and Inference	Final output; raw output; logits; Chain of Thought (summary and raw); refusal patterns	• Jailbreak/safety bypass • Model theft (distillation via systematic querying of outputs/logits to train a replica or parts of a replica) • Capability reconstruction.
Model Internals (Read)	Activations (read); weights (read); architecture states	• Capability reconstruction • Model theft (direct exfiltration of weights enabling exact replication) • Jailbreak/safety bypass • Accidental exposure.
Model Internals (Write)	Activation modification; weights (write/fine-tuning); safety control modification	• Model manipulation • Jailbreak/safety bypass.
Configuration	System prompts; safety instructions; environment parameters	• Jailbreak/safety bypass • Capability reconstruction • Accidental exposure.
Training and Evaluation Data	Training metadata; pre-training samples; post-training data; evaluation data and results; threat models	• Model theft (not that it alone enables it, but access to data may support replication efforts) • Capability reconstruction • Model manipulation (knowledge of training process enables targeted data poisoning) • Accidental exposure • Jailbreak/safety bypass.
Environment and Scaffolding	Computational tools (code execution capabilities); external system access (API, network access); API credentials (authentication tokens); input/output filters	• Credential compromise • Access persistence • Jailbreak/safety bypass.
Compute/ Inference Infrastructure	Inference server architecture; hardware specifications; network specifications; power and cooling specifications	• Model theft • Capability reconstruction (hardware and compute information for reconstruction) • Model manipulation • Jailbreak/safety bypass • Credential compromise • Access persistence.

Source: The authors.

The Access–Risk Matrix

This chapter introduces the Access–Risk Matrix, which addresses how different access arrangements shape the likelihood or severity of the security risks (Table 3). By systematically mapping access modalities against risk categories, the matrix provides a practical tool for AI developers, evaluators and policymakers to assess and mitigate security risks in specific closed-model evaluation contexts and to build on this to expand the mapping of risks as evaluation levels and modalities evolve. The matrix provides a framework for assessing and categorising existing and emergent security risks that have or might arise from access provided for third-party assessments. Table 3 maps seven types of access (detailed in the second chapter) against seven key risk categories (see the third chapter for a description). The cells in between describe the security concerns that emerge from the specific combination of access type and risk category, with citations indicating the strength of evidence supporting each risk.

Through the consolidation of this information into a single reference, the matrix enables model developers, evaluators and policymakers to quickly identify which risks are well-documented versus theoretical and to understand how different types of access create distinct security challenges.

The matrix is designed for critical thinking and decision-making throughout the evaluation access lifecycle. When a third party requests access, developers can use the table as a reference to identify which existing and potential risks are most relevant to the proposed access type and stated evaluation purpose and then design appropriate mitigations targeting those specific concerns. More importantly, this matrix aims to be a living document and framework for the AI safety community to further develop.

Methodology Note

Severity levels reflect how each access type might enable the different classes of security risks. The designations in Table 3 are as follows:

- **High risk (HIGH)**: When access type is a primary mechanism for threat amplification, and without which the attack/compromise is significantly harder to execute.
- **Medium risk (MED)**: When the access type meaningfully amplifies or facilitates the threat but is not the primary mechanism.
- **Low risk (LOW)**: When an access type has an indirect or limited contribution to the threat.

Designations of severity level are informed by academic research, publicly reported incidents and expert consultations conducted through the SAFA-TF.

The designation of risk levels in the matrix below is merely indicative rather than conclusive, and is open to reclassification depending on context. The authors acknowledge that the matrix will benefit from research on access-specific indicators as this research evolves and is publicly shared, thus enabling it to more concretely designate actual access security risks from potential ones within specific organisations.

Table 3: Access–Risk Matrix for Third-Party Evaluator Access to Closed Frontier Models

Exploitation Pathways		Types of Security Risks							
		Structural Exploitation			Adversarial Probing	Outcome	Operational Failures	Cross-cutting Vectors	
Security Risk Categories		Model Theft	Capability Reconstruction	Model Manipulation	Jailbreak/Safety Bypass	Weaponisation	Accidental Exposure	Credential Compromise	Access Persistence
Access Type	Access Level								
Query and Inference	AL1 AL2	MED • Distillation via systematic queries. • API abuse for replication.[60]	LOW • Capability mapping through probing. • Behavioural fingerprinting.[61]	N/A	HIGH • Competing objectives. • Mismatched generalisation. • Refusal pattern exploitation.[62]	MED • Dangerous capability docs. • 'Weaponisable' jailbreak dissemination.[63]	LOW • Unintended disclosure of safety vulnerabilities.	MED • Red team or evaluator credentials leaked, not properly safeguarded or shared inadequately.	MED • Long-term systematic access enabling cumulative and persistent information extraction. • Continuous API access beyond approved scope of access.

60. HM Government, 'AI Insights: Model Distillation', <https://www.gov.uk/government/publications/ai-insights/ai-insights-model-distillation-html>, accessed 13 April 2026; Florian Tramèr et al., 'Stealing Machine Learning Models via Prediction APIs', in 25th USENIX Security Symposium, USENIX Association, Austin, Texas, 10–12 August 2016, pp. 601–18; Carlini et al., 'Stealing Part of a Production Language Model', pp. 5680–705.
61. Dario Pasquini, Evgenios M Kornaropoulos and Giuseppe Ateniese, 'LLMmap: Fingerprinting for Large Language Models', arXiv preprint arXiv:2407.15847, July 2024, <https://arxiv.org/abs/2407.15847>, accessed 13 April 2026.
62. Alexander Wei, Nika Haghtalab and Jacob Steinhardt, 'Jailbroken: How Does LLM Safety Training Fail?', arXiv preprint arXiv:2307.02483, July 2023, <https://arxiv.org/abs/2307.02483>, accessed 13 April 2026; '[C]ompeting objectives occur when a model's pretraining and instruction-following objectives are put at odds with its safety objective.' See p. 2.; '[M]ismatched generalization arises when inputs are out-of-distribution for a model's safety training data but within the scope of its broad pretraining corpus.'
63. Muhammad Salmaan Sid et al., 'Frontier AI Systems Have Surpassed the Self-Replicating Red Line', arXiv preprint arXiv:2504.18565, April 2025, <https://arxiv.org/abs/2504.18565>, accessed 13 April 2026; Unit 42, 'The Dual-Use Dilemma of AI'.

Types of Security Risks									
Exploitation Pathways		**Structural Exploitation**			**Adversarial Probing**	**Outcome**	**Operational Failures**	**Cross-cutting Vectors**	
Security Risk Categories		**Model Theft**	**Capability Reconstruction**	**Model Manipulation**	**Jailbreak/Safety Bypass**	**Weaponisation**	**Accidental Exposure**	**Credential Compromise**	**Access Persistence**
Access Type	**Access Level**								
Model Internals (Read)	AL2 AL3	HIGH • Direct weight/ parameter exfiltration. • Theft of architectural blueprint.[64]	MED • Design pattern extraction supporting other attacks.	N/A	MED • White-box access enables targeted safety bypass and/or internal model inspection.[65]	MED • Insights into architecture weaponised for adversarial attacks. • Unauthorised access by malicious actor to interpretability data supporting model behaviour manipulation.	HIGH • Leak of architectural innovations and IP. • Misconfigured cloud storage (for example) resulting in accidental weight file.	MED • Research environment credentials compromised.	N/A

64. Nevo et al., 'Securing AI Model Weights'.
65. Casper et al., 'Black-Box Access is Insufficient for Rigorous AI Audits'.

Exploitation Pathways		Types of Security Risks							
		Structural Exploitation			Adversarial Probing	Outcome	Operational Failures	Cross-cutting Vectors	
Security Risk Categories		Model Theft	Capability Reconstruction	Model Manipulation	Jailbreak/Safety Bypass	Weaponisation	Accidental Exposure	Credential Compromise	Access Persistence
Access Type	Access Level								
Model Internals (Write)	AL2 AL3	MED • Direct weight/ parameter exfiltration with alignment removed.[66]	N/A	HIGH • Poisoning. • Backdoors. • Safety control removal.[67]	HIGH • Alignment corruption through fine-tuning. • Removal of safety feature. • Insertion of parameter-level bypass and other potential backdoors.[68]	HIGH • Backdoor exploitation to trigger malicious behaviour.[69]	MED • Experimental modifications leading to unintended consequences/ behaviours. • Inadequate parameters for alignment research creating unforeseen risks (emergent capabilities).	MED • Enabling privilege escalation and/or privileged access to malicious actors.	HIGH • Persistent write access enabling ongoing unauthorised modifications.

66. Unit 42, 'The Dual-Use Dilemma of AI'.
67. Evan Hubinger et al., 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training', arXiv preprint arXiv:2401.05566, January 2024, <https://arxiv.org/abs/2401.05566>, accessed 13 April 2026; UK AI Safety Institute, Alan Turing Institute and Anthropic, 'Examining Backdoor Data Poisoning at Scale', 9 October 2025, <https://www.aisi.gov.uk/blog/examining-backdoor-data-poisoning-at-scale>, accessed 23 April 2026; Mithril Security, 'PoisonGPT: How We Poisoned an LLM', 2023, <https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/>, accessed 13 April 2026.
68. Simon Lermen et al., 'LoRA Fine-Tuning Efficiently Undoes Safety Training from Llama 2-Chat 70B', arXiv preprint arXiv:2310.20624, October 2023, <https://arxiv.org/abs/2310.20624>, accessed 13 April 2026.
69. Hubinger et al., 'Sleeper Agents'.

Exploitation Pathways		Types of Security Risks							
		Structural Exploitation			Adversarial Probing	Outcome	Operational Failures	Cross-cutting Vectors	
Security Risk Categories		Model Theft	Capability Reconstruction	Model Manipulation	Jailbreak/Safety Bypass	Weaponisation	Accidental Exposure	Credential Compromise	Access Persistence
Access Type	Access Level								
Configuration	AL1 AL2	N/A	N/A	LOW • Configuration weaknesses creating bypass opportunities.	MED • Configuration weaknesses creating bypass opportunities. • System prompt knowledge enables targeted injection.[70]	N/A	MED • System prompt leakage; Identity Access Management (IAM) miscon-figurations.[71]	MED • Misconfigured or inadequate IAM configurations.	MED • Configuration access enabling persistent infrastructure foothold.
Training Data	AL2 AL3	N/A	HIGH • Methodology, pipelines, alignment techniques are the blueprint.	MED • Targeted data poisoning of training pipeline.[72]	LOW • Privacy mechanism bypass discovery.[73]	MED • Dataset knowledge weaponised for targeted attacks.[74]	HIGH • Inadvertent exposure of sensitive training data.	MED • Data storage/ pipeline credentials compromised.	N/A
Evaluation Data and Results	AL1 AL2	N/A	MED • Reveals capability landscape and development trajectory.	LOW • Reveals capability landscape and development trajectory.	MED • Evaluation results reveal known vulnerabilities and defences.	LOW • Threat models could inform attacker targeting.	HIGH • Sensitive findings and methodologies exposed.	MED • Evaluation infrastructure credentials compromised.	N/A

70. OWASP, 'Top 10 for Large Language Model Applications 2025', 2025, <https://genai.owasp.org/>, accessed 13 April 2026. See: LLM07: System Prompt Leakage; LLM01: Prompt Injection.
71. *Ibid.*
72. UK AI Safety Institute, Alan Turing Institute and Anthropic, 'Examining Backdoor Data Poisoning at Scale'.
73. Milad Nasr et al., 'Scalable Extraction of Training Data from (Production) Language Models', arXiv preprint arXiv:2311.17035, November 2023, <https://arxiv.org/abs/2311.17035>, accessed 13 April 2026.
74. Nicholas Carlini and Andreas Terzis, 'Poisoning and Backdooring Contrastive Learning', arXiv preprint arXiv:2106.09667, June 2021, <https://arxiv.org/abs/2106.09667>, accessed 13 April 2026.

Exploitation Pathways		Types of Security Risks							
		Structural Exploitation			Adversarial Probing	Outcome	Operational Failures	Cross-cutting Vectors	
Security Risk Categories		Model Theft	Capability Reconstruction	Model Manipulation	Jailbreak/Safety Bypass	Weaponisation	Accidental Exposure	Credential Compromise	Access Persistence
Access Type	Access Level								
Environment and Scaffolding	AL1 AL2	N/A	N/A	N/A	MED • External content processed via tools enables indirect prompt injection.[75]	N/A	N/A	N/A	HIGH • Backdoors in tools connected to models (for example, in an agentic AI architecture) could grant persistent access to attackers.
Compute/ Inference Infrastructure	AL2 AL3	MED • Side-channel attacks on inference infrastructure.	LOW • Hardware/ compute requirements revealed.	MED • Serving-layer tampering alters model behaviour.	LOW • Infrastructure exploits bypass app-level safety controls.	MED • Hijacked compute for harmful inference at scale.	MED • Misconfigured infrastructure exposing endpoints.	HIGH • Infrastructure credentials. • High-value targets.	HIGH • Infrastructure access hardest to detect and revoke.

Source: The authors.

75. Reddy and Gujral, 'EchoLeak'.

Key Patterns

At least four key patterns and related lessons emerge from the Access–Risk Matrix:

1. All access types come with some level of risk. This ranges from the black-box access (AL1), which is seen to potentially enable some form of model distillation, to the assumed high-risk white-box access (AL3). This shows that security risk mitigation efforts should be rigorously applied across all access types (AL1–AL3). As seen above, even in cases of low-severity-level access types, malicious actors might use those for reconnaissance to gain initial access to a system.
2. Write access to model internals represents the access type with the highest level of risk, given that it provides the adversary with the capacity to potentially tamper with model behaviour directly when compared to read-only access types. By contrast, query and inference, and evaluation data and results, represent the ones with the lowest (overall) risk level if compared to others.
3. Cross-cutting vectors such as credential compromise and access persistence are not only part of the chain of action conducted by threat actors, but also crucially amplify threats across every access type and level, making access revocation/management and credential management core mitigations regardless of the type of evaluation being conducted.
4. Training and evaluation data provide insights into different potential threat activity pathways and are one example of the subjectivity of the risk levels. The former includes indicators of how the model was built (datasets used, methodology and alignment techniques) while the latter can indicate what the model can and cannot do (red-teaming findings, benchmark results and vulnerability assessments). In this regard, training data could support adversaries in capability reconstruction, while evaluation data could enable jailbreak/safety bypass. Even if the latter is less sensitive, if an adversary gathers, for example, information about the model's threat model, it is nonetheless an enabler for further attacks.

These findings reinforce the paper's central argument: a shared, risk-informed framework for evaluator access is essential to enabling meaningful third-party assessments while proportionately managing security risks.

Security Controls to Strengthen Access

Recommendations are outlined in Table 4 for mitigating security risks, based on existing research (mostly focused on the security of closed models and model weights, as referred to in the previous chapters) and input from SAFA-TF members.

The controls in Table 4 are designed for a context where:

1. Access is deliberately granted to third parties for legitimate evaluation purposes.
2. Access is determined by contractual obligations and terms or a Memorandum of Understanding.
3. Unintended threats also arise from misconfigurations or a lack of proper controls by evaluators or developers.

Security controls are organised by risk category and access types as outlined in the Access–Risk Matrix the previous chapter.

The controls in Table 4 are also in line with existing principles already used in other fields, such as cybersecurity, data security and protection, and risk management (for example, least privilege, assume breach, need-to-know, data minimisation, time-bound access, proportionality, and transparency and accountability).

Table 4: Security Mitigation Actions for Third-Party Access

	Technical Controls	Procedural Controls	Contractual Controls
Model Theft	Output watermarking; security enclaves/Trusted Execution Environments (TEEs) for weight access; confidential inference systems; no-export environments RESPONSIBILITY: **D**	Evaluation-scoped access; session logging and review; evaluators should have policies for secure storage, transit and retention of completions/transcripts; limiting API access; security screening of personnel RESPONSIBILITY: **J**	IP protection clauses; data destruction requirements RESPONSIBILITY: **J**
Capability Reconstruction	Compartmentalised architecture; differential privacy for data; scoped API endpoints (tailored to defined evaluation task) RESPONSIBILITY: **D**	Need-to-know access approval; query pattern review RESPONSIBILITY: **D**	IP non-compete clauses; strict prohibition on 'distillation-style' data logging during the evaluation period RESPONSIBILITY: **J**
Jailbreak	Sandboxed fine-tuning; restricted access to reinforcement learning from human feedback components RESPONSIBILITY: **D**	Coordinated disclosure protocols; time-limited red-team access window RESPONSIBILITY: **J**	Responsible disclosure agreements RESPONSIBILITY: **J**
Weaponisation	Tiered access to evaluators depending on clearance or scope of evaluation RESPONSIBILITY: **D**	Third-party vetting; structured and coordinated disclosure for dangerous capabilities RESPONSIBILITY: **J**	Mandatory disclosure of 'dual-use' capabilities to the developer within a pre-defined number of hours of discovery; liability waivers/ safe harbour provisions for good faith research RESPONSIBILITY: **J**
Accidental Exposure	Security enclaves/TEEs of data-loss prevention tools; secure handling environments RESPONSIBILITY: **D**	Security training; protocols for unexpected discoveries; develop security standard for third-party access RESPONSIBILITY: **J**	Incident notification requirements/ agreement; joint security standards for evaluation RESPONSIBILITY: **J**
Credential Compromise	Multifactor authentication; hardware security keys; principle of least privilege RESPONSIBILITY: **D**	Security training; anomaly monitoring; develop security standard for third-party access RESPONSIBILITY: **J**	Notification of breach requirements; joint security standards for evaluation; mandatory 'security posture' assessment for evaluators before access is granted RESPONSIBILITY: **J**
Access Persistence	Programmed and automatic credential expiration RESPONSIBILITY: **D**	Regular access verification/ monitoring; termination checklists/ monitor; offboarding procedures RESPONSIBILITY: **J**	Time-bound/access expiration clauses; termination procedures; identification of dormant evaluation accounts RESPONSIBILITY: **J**

Source: The authors.

Legend
D = DEVELOPER
J = JOINT

The Way Forward: Towards a Shared Governance Framework

As the preceding chapters have shown, the security risks associated with third-party access to frontier AI models are not uniform or static. They vary by access type, evolve with the threat landscape and cut across technical, procedural and institutional boundaries. How the ecosystem navigates these risks will shape not only the viability of independent evaluation but also the broader architecture of accountability in AI governance.

This paper has:

- mapped the security risks associated with providing third-party evaluators with access to frontier AI models;
- proposed a taxonomy of seven access types and security risk categories;
- presented an Access–Risk Matrix to reflect existing and potential threats according to the degrees of access;
- outlined security controls for each access type.

These contributions are a starting point of a multistakeholder effort to develop a shared understanding of how to adequately support access, which can enable further innovation in AI models and applications in a safe and secure manner. Translating these contributions into practice will require a shared effort, that is, a shared governance framework that is collaboratively developed by evaluators, AI developers and policymakers – one that treats secure access not as a constraint on evaluation or innovation but as the infrastructure that enables it.

At the time of writing, there was no international shared standard or framework for providing third parties with access to frontier AI models. Organisations such as OpenAI have, for example, launched a 'Trusted Access for Cyber' programme, which operates as 'an identity and trust-based framework designed to help ensure enhanced

cyber capabilities are placed in the right hands'.[76] The programme provides a vetting system to enable cybersecurity researchers to use model versions that are 'permissive models' and can 'accelerate legitimate defensive work'. However, it is cybersecurity-specific and not inclusive of other types of evaluations and tests.

Access decisions remain ad hoc, security expectations are inconsistent and the language used to describe access levels varies across jurisdictions, organisations and agreements. The SAFA-TF has identified several topics of concern, and the section below draws from those discussions to propose coordinated progress across three pillars.

Pillar 1: Harmonising Language and Access Tiers

Clarify and define 'adequate access' to ensure that disparate discussions surrounding third-party access and evaluations use like-for-like language.

Currently, understandings of minimum or adequate access are disparate and highly dependent on a particular developer's practices, evaluators' individual methodologies and other bilaterally set agreements. Different evaluations and intents of access will indeed require different levels of access, which makes a single, all-encompassing definition of adequate access challenging, but not impossible.

While defining minimum or adequate access to a frontier AI model may be a point of contention between AI developers and evaluators, the level of access granted to third-party evaluators is a significant factor in assessing the robustness and value of those evaluations. Reduced or restricted access during an assessment can curtail evaluators' ability to identify potential risks and evaluate the frontier models' safeguards.

Potential actions towards harmonisation could include:

- **Adopting a shared taxonomy of access levels.** The AL1/AL2/AL3 framework proposed by Jacob Charnock and others, and used throughout this paper, offers one avenue for broadly representing the depth of access required.[77] Whether this or an alternative framework prevails, the ecosystem of stakeholders would benefit from converging on a common set of terms so that developers, evaluators and policymakers are working from the same reference points when discussing what access is being granted, to whom and under what conditions. This should come hand in hand with context-specific adjustments to granting access, depending on sensitivity, timing, scale and other factors, to tailor the required depth of access.

76. OpenAI, 'Introducing Trusted Access for Cyber', 5 February 2026, <https://openai.com/index/trusted-access-for-cyber/>, accessed 13 April 2026.
77. Charnock et al., 'Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations'.

- **Mapping access requirements to evaluation objectives.** Different evaluation types (for example, pre- and post-deployment) require different levels and types of access. A shared framework should continue to make these correspondences explicit – as with the Access–Risk Matrix – thereby reducing ambiguity about what access is appropriate for a given evaluation scope.
- **Standardising definitions for terms and concepts related to access and evaluations.** As the EU AI Act's CoP, UK AI Safety Institute frameworks and other international safety commitments develop in parallel, there is a risk of fragmentation where adequate access ends up being defined in different ways across multiple jurisdictions. Early coordination between these processes could help ensure interoperability and scalability and/or further specialisation of evaluator services, business models and research development.

Pillar 2: Operationalising Secure Access

Taskforce members emphasised the need for developers and AI companies to develop clearer and more standardised access frameworks, which can then be tailored to the various access requirements while ensuring an adequate security and safety baseline.

Facilitating the development of standardised and recognised secure access frameworks requires maintaining and supporting partnership forums dedicated to supporting secure access. To ensure balanced perspectives and potential sponsor or evaluation-specific requirements, all initiatives should encompass representatives from across the AI developers, deployers and evaluators from the assurance sector. Though not due to regulatory requirements at present, conveners should consider including regulators and government governance specialists in relevant discussions. Fortunately, a range of forums exist, encompassing evaluation-specific organisations and wider fora, including the Frontier Model Forum, the RUSI SAFA-TF and the newly founded AI Evaluator Forum (along with groups such as AVERI, among other non-evaluation-specific initiatives led by the AISI, OECD AI Observatory and other convening and research bodies.

A common vocabulary is necessary but not sufficient. The ecosystem also needs shared practices, that is, agreed-upon standards for how access is granted, secured, monitored and revoked. The security controls outlined in the fifth chapter represent an initial contribution to this effort, but operationalising them requires action from multiple parties.

Areas where shared standards and best practices could be developed include:

- **Data minimisation and purpose limitation.** As previously indicated in this paper, evaluators should request only the access necessary for their specific evaluation scope, and developers should design access environments that make it technically difficult to extract information beyond what is needed, while making sure adequate access is provided for evaluation. [78] Both parties share an interest in limiting the attack surface without limiting the evaluation surface.
- **Bridging cybersecurity and AI evaluation expertise.** Many of the risks identified in this paper, such as credential compromise, access persistence and data exfiltration, are well understood by cybersecurity experts. The AI evaluation community should actively engage with cybersecurity professionals and vice versa. This cross-pollination is key not only to sharing a vocabulary but also to learning about existing or emerging cybersecurity standards that can be adapted to support AI discussions on the access–risk nexus. AI evaluators bring domain expertise in model behaviour, while cybersecurity professionals bring decades of experience in managing the access risks identified in this paper. Structured dialogue between these communities through joint workshops or conferences would help accelerate the development of additional security practices.
- **Incident response and coordinated disclosure.** One of the elements raised in the mitigations section the fifth chapter of the paper included incident notification and vulnerability reporting. Given the joint responsibility (developers and evaluators) for frontier model reporting, both would benefit from further dialogue on incident response procedures for security breaches, accidental exposure of sensitive information, and the discovery of dangerous capabilities during evaluation. Coordinated disclosure protocols of vulnerabilities and bugs are already common in cybersecurity research and could be further formalised for the AI evaluation context, building on the responsible disclosure agreements outlined in Table 4.

Pillar 3: Building Feedback Loops for Continuous Improvement

The threat landscape for frontier AI evaluation is not static. New model architectures, new evaluation methodologies and new access patterns – particularly as agentic AI systems and multi-model deployments become more prevalent – will introduce risks not fully captured by the current taxonomy. A governance framework must therefore be designed to learn and adapt.

78. Reuel et al., 'Open Problems in Technical AI Governance'.

Mechanisms for continuous improvement could include:

- **Empirical validation of the risk framework.** The Access–Risk Matrix and severity ratings proposed in this paper are based on expert assessment and available evidence. As more evaluations are conducted under structured access agreements, real-world data on threat materialisation, control effectiveness and threat types should be systematically collected and used to refine the matrix.
- **Coordinated incident learning and information sharing to improve access-related security.** AI labs have recently begun to publish their own version of cyber threat intelligence reports. As they continue to develop the model, periodicity and frequency with which they publicly disclose information, it would be helpful to explore public/selective/private publication of access-specific indicators of threats across industry. This could be facilitated by existing cross-industry initiatives but should preferably be a multistakeholder coalition. Information sharing about security incidents – whether arising from third-party evaluation access or from the wider AI development process – represents valuable data for improving access governance. While there are multiple existing discussions on AI incident reporting, the proposed action calls for trialling information sharing and framework development focused on strengthening access. A clear, specific and tailored objective such as this could serve as a case study or pilot for building trust in information sharing across industry. Models such as the Global Internet Forum to Counter Terrorism is one such example where technical and governance measures have helped coordinate industry action, standardisation and even external dialogue with governments.[79] Information sharing should be the primary objective, followed by the designation or establishment of the required platform/model for the dialogue and the stakeholders required for adequate cross-stakeholder action.

79. Naureen Chowdhury Fink and Erin Saltman, 'Fighting Terror with Tech: The Evolution of the Global Internet Forum to Counter Terrorism', in Maia Levy Daniel et al. (eds), *Trust, Safety, and the Internet We Share: Multistakeholder Insight*s (Abingdon: Taylor & Francis, 2026).

- **Periodic review and iteration of the Access–Risk Matrix.** As previously noted in this paper, the matrix should be treated as a living document, subject to periodic review as the evidence base grows. Future iterations should draw on a broader evidence base, including structured threat modelling exercises conducted jointly by developers, evaluators and cybersecurity experts.
- **International coordination.** The security risks of third-party access to frontier models are not confined to any single jurisdiction. Frameworks developed in the EU, the UK, the US and elsewhere should be designed with interoperability in mind so that evaluators operating across borders are not subject to contradictory requirements, and so that insights gained in one context can inform practice elsewhere.

About the Authors

Louise Marie Hurel is a Senior Research Fellow in the Cyber and Tech research group at RUSI. Her research interests include incident response, cyber capacity building, cyber diplomacy and non-governmental actors' engagement in cyber security. Louise Marie completed her PhD in Data, Networks and Society at the London School of Economics' (LSE) Department of Media and Communications. She has an MSc in Media and Communications (Data and Society) from the LSE and a BA in International Relations from the Pontifical Catholic University of Rio de Janeiro.

She is an advisory board member of the Global Forum of Cyber Expertise, Carnegie Endowment's Partnership for Countering Influence Operations' and the Centre for Information Resilience.

Elijah Glantz is a Research Fellow in the Organised Crime and Policing research group (OCP) at RUSI. His research at RUSI focuses on criminal networks structures, illicit finance and law enforcement responses around the world. Specific projects include illicit finance and law enforcement capacity in sub-Saharan Africa, state involvement in organised crime and narcotics, and trends in UK organised acquisitive crime. He is the Co-Project Manager and Lead Editor for the Strategic Hub for Organised Crime Research (SHOC) – OCP's blog for practitioners, academics and experts on organised crime. Elijah has an MSc (Distinction) in Conflict Studies from the London School of Economics and a BA in Political Humanities from Sciences Po Paris.

Daniel Cuthbert is a senior security researcher, technologist and long-standing contributor to the global cybersecurity community. He serves as Global Head of Cyber Security Research for the Santander Group, where he focuses on advancing detection engineering, cryptographic safety and applied security research across complex financial systems. As a founding member of The Open Web Application Security Project (OWASP), where he was co-author of the OWASP Testing Guide and the OWASP Application Security Verification Standard (ASVS), Daniel has played a defining role in shaping modern approaches to secure software development. His work continues to influence standards and practitioners globally.

195 years of independent thinking on defence and security

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

Since its foundation in 1831, RUSI has relied on its members to support its activities. Together with revenue from research, publications and conferences, RUSI has sustained its political independence for 195 years.

RUSI is registered as a charity in England and Wales

Charity number: 210639

VAT number: GB752275038