5b. Are there effective ways to create safeguards around foundation models, to ensure that model weights do not become available?

Executive Summary

To ensure the security of foundation model weights, organizations should implement a comprehensive set of policies and procedures:

- Adopt intrusion detection systems (IDS) aligned with ISO/IEC 27001 and 27039 standards
- Conduct regular audits as per ISO/IEC 19011
- Perform Data Protection Impact Assessments (DPIA) in accordance with ISO/IEC 29134
- Mandate the use of secure execution environments
- Require certified intrusion detection systems
- Implement API security measures
- Use advanced encryption techniques for model weights
- Establish procedures for documenting data access and usage
- Segment model information across secure environments
- Implement a structured audit schedule with a feedback loop for prompt issue resolution

By enacting these policies and procedures, organizations can significantly enhance the security of their foundation models and protect against unauthorized access to model weights.

For the scope of this RFC, we shall assume that the safeguards are being discussed in the context of the security of the party that conducted the initial training run of the model (e.g. OpenAI GPT4, Anthropic Claude 3, DeepMind Gemini, etc.)

Notably, OpenAI [1] and Anthropic [2] have both acknowledged the risks from the lack of adequate cybersecurity of model weights. Security measures such as standard security infrastructure, compliance drills and mandatory external reviews have been mentioned.

Recommended Safeguards

Intrusion Detection Systems

Labs should ensure that IDS is seamlessly integrated into their existing information security management system (ISMS), as outlined in ISO/IEC 27001 [3], with specific attention to the recommendations provided by ISO/IEC 27039 [4] for intrusion detection and prevention.

Given the unique challenges in protecting foundation models [5], it is essential for labs to customize and configure their IDS solutions to detect specific types of intrusions that could lead to unauthorized access to model weights. This involves setting up anomaly detection systems that can identify unusual patterns of access or attempts to exploit vulnerabilities specifically related to the storage and transmission of model weights.

Regular Auditing Schedules

Implement routine audits in line with the ISO/IEC 19011 [6] standards to assess the security of the training environment, the efficacy of encryption technologies, and adherence to data handling protocols. These audits should also evaluate the access controls and check for any unauthorized attempts to access the model weights and biases. Ensures continuous vigilance and early detection of potential security lapses, maintaining a high-security standard throughout the training process.

Audits must record conformities and non-conformities.

Systematic DPIA Procedures

Establish a structured process for conducting DPIAs, including clear guidelines on when and how to carry out these assessments. Conduct Data Protection Impact Assessments (DPIA) for new AI projects as per ISO/IEC 29134 [7].

The DPIA should include Risk evaluation, compliance analysis, consequences and their level of impact. Proper stakeholder consultation across teams has to be carried out.

Implement strong encryption and access control measures as per ISO/IEC 27001 standards [8].

Utilization of Secure Execution Environments

- Deploy models within secure and verified platforms or containers that offer robust security features, such as isolation from other network processes.
- Regularly update and patch these environments to address new security vulnerabilities.

Certification of Intrusion Detection Systems

Adopt a policy requiring that all intrusion detection systems used for monitoring AI models be certified against recognised security standards. This ensures a baseline quality and effectiveness in detecting unauthorized access attempts or anomalous behavior indicative of a security breach.

Incorporation of API Security Measures

- Implement API gateways with robust authentication and authorization checks.
- Use rate limiting, encryption, and regular scanning for vulnerabilities in APIs.
- Train staff on API security best practices, including how to identify and mitigate potential threats.

Model Encryption Standard Adoption and Maintenance

- Establish a standard procedure for encrypting AI models before deployment, ensuring data integrity and confidentiality.
- Regularly review and update encryption methods to ensure they remain effective against new decryption techniques and threats.

Documentation and Review

Implement a documentation system to record the DPIA process, findings, and actions taken. Regularly review and update DPIA methodologies to reflect changes in regulations, technologies, or organizational practices.

Documentation of Data Access and Usage

Maintain detailed logs of who accesses the model data, when, and for what purpose. This should include tracking the movement of training datasets, logging any changes to model weights, and documenting any external data inputs. Provides a clear audit trail that can be crucial for tracing the source of any leak or un-authorised access, enhancing accountability and transparency in the model training phase.

Enhanced Model Encryption and Access Control

- Policy: Implement advanced encryption techniques specifically for AI model weights, biases, and architecture details.
- Workflow Change: Only authorized personnel can decrypt and access the full details of the model. Access logs must be maintained for auditing.

Segmentation of Model Information

- Policy: Segregate model information (weights, biases, architecture) across different secure environments.
- Workflow Change: Researchers access only the segments of the model necessary for their work, preventing any single individual from accessing the entire model's details.

API Security Updates and Monitoring Systems

- Implement an ongoing process for updating API security measures, including patch management and monitoring for unusual access patterns or potential breaches.
- Deploy monitoring systems that can provide real-time alerts on potential security incidents affecting API integrity.

Structured Audit Schedule and Feedback Implementation

- Develop a structured schedule for regular security audits, ensuring consistent evaluation of the deployed AI system's security.
- Create a feedback loop where findings from audits are promptly addressed and necessary updates or changes are implemented in the deployment strategy.

References

[1] ISO/IEC 27001 Information Security Management

[2] ISO/IEC 29134:2023 Guidelines for privacy impact assessment

[3] ISO/IEC 27039:2015 Information technology — Security techniques — Selection, deployment and operations of intrusion detection and prevention systems (IDPS)

[4] arXiv:2108.07258 [cs.CR] - Challenges in Securing the Future of Artificial Intelligence

[5] ISO 19011:2018 Guidelines for Auditing Management Systems

[6] ISO/IEC 27001 Information Security Management

[7] Anthropic Response - Internal AI safety policy and securing model weights

[8] OpenAI Preparedness Framework