Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/AN16.20260018
Published: June 23, 2026

Reducing Model Exposure Risk Through AI-Governed Feature Access Control

Chhaya Gunawat
System Development Engineer
Amazon, California, United States
chhayagunawat@gmail.com

Jay Sunil Nankani
Software Developer, Redfin
California, United States
jay.nankani1@gmail.com

Prathana Gunawat
ML Cloud Engineer
Amazon, Seattle, United States
prathanagunawat98@gmail.com

Abstract

In recent years, the incorporation of machine learning models within enterprise and cloud computing applications has introduced significant risks of model exposure from unauthorized feature access, abusive inferences, and other forms of exploitation. Current methods of access control rely mainly on role-based policies that are not dynamic enough to cater to emerging threats, changing user behavior, and context-dependent security policies. In this paper, we propose an AI-powered access control scheme for managing feature-level exposure in machine learning models and reducing the risk of model exposure to a minimum. Our approach leverages behavioral analysis and risk scoring capabilities to enforce access restrictions based on contextual analysis of feature-level usage and activities. In particular, our policy engine analyzes the interaction between users and model features, detects suspicious activity, and restricts access accordingly. Experiments carried out in a closed environment have shown that the proposed framework effectively limits unauthorized feature exposure while preserving the operational efficiency of deployed models.

Keywords
Reinforcement Learning, Autonomous incident response, Threat detection.

1. Introduction
The extensive use of machine learning (ML) models in cloud computing, edge computing, and enterprise applications has raised serious worries about the risks associated with exposure and misuse of the models. Modern ML models are accessible via APIs and services and thus are open to attacks like model stealing, inference attacks, data leakage, and feature abuse attacks. As models handle more and more critical decisions and sensitive data, controlling access to model features becomes an essential task in implementing artificial intelligence systems responsibly and securely.

The traditional methods of access control and authentication implemented in ML systems usually include static access control rules and role-based authentication and perimeter security. Although such techniques work well with ordinary software applications, they are inadequate when dealing with the threats posed to machine learning models, which

change depending on the user and query behavior. This makes attackers able to take advantage of these weaknesses and try out feature abuse again and again until the model features get revealed.

Recent researches have shown that the risk of exposure to models is not just about the precision or architecture of the model but is significantly impacted by the way in which features and inference abilities are managed. When unrestricted access is available to features, it provides adversaries with enough opportunities to probe systematically and ultimately succeed in attacking by means of model inversion, membership inference, and feature extraction. There is an urgent need for mechanisms that manage risks at the level of features rather than the model itself or the users.
In order to tackle the aforementioned issues, this paper presents an innovative approach where the access to the features of machine learning models is controlled by artificial intelligence. The process involves analyzing access requests in real-time in order to identify risk factors and enforce appropriate measures accordingly.

The major contributions of this study can be summarized as three: (i) the recognition of feature-level access as a critical attack surface for models' exposure risks, (ii) the development of an AI-based governance model that can automatically enforce access controls based on behavioral risk assessments, and (iii) a practical evaluation that proves the efficiency of the proposed method in minimizing feature exposure while preserving model accuracy. The rest of the paper is organized as follows. In Section II, we review related studies; in Section III, we introduce our model; in Section IV, we conduct experiments and provide analyses; and finally, in Section V, we conclude our findings and suggest future research topics.

1.1 Traditional Solutions

Classical means of protecting machine learning algorithms from potential threats are mostly grounded in traditional approaches to software protection and information system governance. As such, they are mostly focused on perimeter defense, access control policies that are generally static, and authorization techniques that are quite broad and not tailored specifically for dealing with machine learning models' vulnerabilities.

One of the most popular techniques that can be used to protect machine learning models is the Role-Based Access Control (RBAC). In this technique, users have access to certain services depending on their role (developer, administrator, or user). RBAC provides a simple yet effective access control method, but, at the same time, it is unable to deal with dynamic security threats since users who have once been given access to certain resources will be able to use them without any limitations even after exhibiting malicious behavior. Thus, RBAC fails to counteract inference attacks.

API authentication and authorization techniques represent another well-known approach to securing ML models. These techniques include authentication using API keys or OAuth, as well as the use of a security gateway for managing requests. Attackers who obtain valid credentials can systematically probe model endpoints, extract sensitive feature relationships, or reverse-engineer model parameters without triggering conventional security alarms.

In addition to the use of application-based security controls, network-based measures such as firewalls, IDSs, and IPSs are widely used to safeguard the ML system from any potential attacks. However, these technologies are only good at detecting and preventing abnormal network activity or any attempts of unauthorized connection but cannot be used to detect any semantic interaction with the machine learning model. As a result, malicious requests to run inferences usually look legitimate from the network activity point of view.

Some companies rely on the use of rate limiting and usage quotas as a preventive measure against automated attacks. Rate limiting can help slow down any attempts of performing brute force attacks on the system but will not make the problem disappear completely. The adversaries would just have enough time to extract the required information slowly but steadily during a considerable amount of time.

To summarize, all the mentioned traditional techniques treat ML models like regular software services and do not account for the specific risks associated with exposing them. These techniques are too rigid, and they do not allow controlling any features or actions related to model exposure.

1.2 Modern Solutions

The evolution of machine learning security has led to more intelligent and adaptable measures to counteract model exposure threats. Contrary to the perimeter-centric security paradigm, current solutions prioritize behavior awareness,

control granularity, and risk assessment in safeguarding ML models against misusage, extraction, and inference attacks.

One prevalent contemporary technique is the Attribute-Based Access Control (ABAC) approach, which considers user identity, request type, date, time, geographical location, and usage behavior when evaluating access rights. Although ABAC enhances scalability than RBAC by introducing conditional access policies, it still involves predetermined access policies and cannot make self-regulated decisions. In the face of sophisticated and advanced attacks, conventional attribute definitions may become irrelevant in detecting and stopping malicious actions.
Differential Privacy (DP) has been considered an extensively explored approach to mitigate information leakage from machine learning models. The process of applying controlled noise in the output of ML models minimizes the chances of inferring sensitive data. Even though differential privacy effectively hides personal data, excessive noise in model results might compromise the quality of predictions. Differential privacy is powerless in defending against attacks that aim at extracting the model's internal mechanism through query repetition.

One such solution is secure model deployment through TEEs or secure enclaves. Hardware-based security measures help secure parameters and execution from any form of direct scrutiny. However, although TEEs are successful at the infrastructure level, they do not protect against logical exposure through application programming interfaces, which enables attackers to infer internal behavior based on outputs alone.

Another solution that has been gaining popularity for defense against model stealing includes adversarial behavior detection or anomaly monitoring. Such systems use advanced statistical and machine learning analysis to detect abnormal query patterns indicative of attacks on ML models. Detection-based systems help reveal useful insights about potential vulnerabilities but lack a preventive approach and may not be proactive enough.

Recent systems include dynamic throttling mechanisms that adjust access limits based on certain risk levels. Throttling systems allow an element of adaptability but only up to the request/user level. They are not capable of controlling model exposure at the attribute/feature level.

In general, contemporary approaches have greatly outperformed the traditional ones since they leverage context awareness and behavior analysis. However, they still have shortcomings in terms of coarse grain control, late reaction, or balancing security and performance. This calls for the development of AI-driven feature access control that combines timely decision-making, fine-grained control at the feature level, and continuous risk evaluation to reduce risks while ensuring usability.

2. The Business Need

With AI systems increasingly becoming key assets for businesses, companies have the challenge of balancing innovation, accessibility, and security. AI models have moved from being an experiment to being tangible assets with monetary worth. The consequence of unregulated access becomes a threat to businesses because it entails risks that affect financial, legal, regulatory, and even reputational standing.

From a business point of view, AI models are intellectual property whose development takes time and requires financial resources. Accessing the model or using reverse engineering can result in a loss of competitive advantage in industries like finance, health care, e-commerce, and cybersecurity. Model access control systems cannot provide any protection for models in situations where they are accessed by all users.

On top of that, regulations such as GDPR and HIPAA impose requirements on how AI models should operate, including data management and transparent decision-making processes. Inappropriate feature exposure could lead to revealing confidential or regulated features using inference attacks, leaving businesses subject to fines. Businesses thus need a way of controlling access to certain features dynamically based on the user's role and intent.

Scalability is another significant challenge. Today, businesses expose their AI models via APIs, making them accessible to internal and external parties including other business units within the same company, external partners, and independent third-party developers. These consumers have different confidence levels. Equal feature access for everyone will increase risks, whereas having several different models will increase operational costs. Businesses therefore need an alternative control mechanism that can be used centrally but still allow differentiation in feature access without having multiple copies of the model.

When it comes to security policy, it would not be enough to apply the same static policy to an ever-changing business environment. In dynamic business settings, user activity varies as well as potential threats. Thus, a business-oriented approach to security needs to be dynamic, allowing for restricting access for malicious activities while at the same time avoiding disrupting legitimate business.

Third, the issue of customer trust and corporate reputation is directly influenced by the inability of the organization to safeguard its AI technologies. Any sort of data breach, AI-biased output, or even the exposure of AI proprietary logic can lead to a loss of confidence in AI-driven technology. It is now understood that the responsible use of AI technology is both a necessity and a business strategy. The application of feature-level access control helps to secure the value of these AI systems from a business perspective. In conclusion, the need for the application of AI-driven feature-level access control stems from the fact that there are many business drivers that make the development of intelligent feature-level access control critical to business success.

3. Proposed Solutions

In response to these emerging threats, the present paper introduces an AG-FAC approach to managing the risks associated with unregulated exposure of machine learning models to third parties. The novel scheme offers dynamic access controls for ML model features based on contextual risk assessments, user trustworthiness, and policy constraints. This contrasts with existing solutions where access control operates either at the level of the entire ML model or the API layer, while the proposed mechanism ensures governance of ML model features with respect to their sensitivity and risk levels.

The AG-FAC system consists of four tightly-coupled modules, namely: (i) Feature Sensitivity Classification; (ii) AI-Driven Risk Assessment; (iii) Policy-Aware Access Enforcement; and (iv) Continuous Governance and Adaptation.

3.1 Feature Sensitivity Classification Layer

The first component of the proposed solution focuses on identifying and classifying model features based on their **exposure risk and business sensitivity**. Not all features contribute equally to model vulnerability; some features may reveal proprietary logic, sensitive user data, or allow adversaries to infer model parameters.
Each feature is assigned a **sensitivity score** derived from:

- **Data criticality** (e.g., personal, financial, or regulated attributes),
- **Model dependency strength** (feature importance or contribution to output),
- **Inference leakage potential** (susceptibility to reconstruction or membership inference),
- **Business impact** (strategic value and IP relevance).

This classification process combines explainable AI techniques such as SHAP or gradient-based attribution with rule-based domain knowledge defined by compliance and business stakeholders. Features are grouped into tiers (e.g., low, medium, high sensitivity), forming the foundation for controlled exposure decisions.

3.2 AI-Driven Risk Assessment Engine

At the core of the proposed framework lies an **AI-driven risk assessment engine** responsible for evaluating exposure risk in real time. Instead of relying on static thresholds, this engine continuously analyzes operational context to determine the appropriate level of feature access.
The risk assessment model considers multiple dynamic inputs, including:

- **User identity and trust level** (internal staff, partners, public consumers),
- **Access behavior patterns** (query frequency, anomaly detection),
- **Request context** (geolocation, device fingerprint, time of access),

- **Historical misuse signals** (previous policy violations or abnormal outputs).

Using supervised and reinforcement learning techniques, the engine learns optimal access decisions by balancing **utility maximization** against **exposure minimization**. When elevated risk is detected, the system can automatically restrict or obfuscate sensitive features while maintaining acceptable model performance.

5.3 Policy-Aware Feature Access Enforcement

The third layer implements **policy-aware enforcement mechanisms** that translate risk scores into actionable access controls. Business, legal, and compliance policies are formalized into machine-readable rules that define allowable feature exposure under different risk conditions.
Enforcement actions include:

- **Feature masking or removal** for unauthorized requests,
- **Feature quantization or noise injection** to reduce information leakage,
- **Adaptive output abstraction** to prevent reverse engineering,
- **Dynamic rate limiting** for high-risk access patterns.

Unlike static rule engines, enforcement decisions are **risk-adaptive** and context-sensitive, allowing legitimate users to retain sufficient model functionality while minimizing exposure to adversarial behavior. This ensures that security controls remain aligned with operational and business objectives.

3.4 Continuous Governance and Adaptive Learning

The proposed solution incorporates a **continuous governance loop** that enables the system to evolve alongside changing threat landscapes and business requirements. All access decisions, risk scores, and enforcement actions are logged and fed back into the learning pipeline.
Key governance functions include:

- **Continuous policy refinement** based on observed outcomes,
- **Model drift detection** to identify emerging exposure risks,
- **Compliance auditing and traceability** for regulatory reporting,
- **Human-in-the-loop oversight** for high-impact decisions.

This adaptive learning mechanism ensures that feature access policies remain effective over time without requiring frequent manual reconfiguration. Governance dashboards provide stakeholders with visibility into exposure risks, enforcement effectiveness, and compliance posture.

3.5 Business and Operational Benefits

From an organizational perspective, the proposed AG-FAC solution delivers multiple tangible benefits:

- **Reduced intellectual property leakage** through feature-level protection,
- **Lower compliance risk** by enforcing regulatory constraints dynamically,
- **Operational scalability** by avoiding model duplication across access tiers,
- **Improved trust and transparency** in AI service delivery.

By integrating AI-driven governance directly into the model access pipeline, organizations can safely expose AI capabilities to diverse user groups while preserving both performance and strategic value.

3.6 Summary of the Proposed Approach

In summary, the proposed AI-Governed Feature Access Control framework moves beyond traditional model-level security by introducing intelligent, adaptive, and business-aligned feature governance. Through the combination of sensitivity classification, real-time risk assessment, policy-aware enforcement, and continuous learning, the solution effectively reduces model exposure risk while enabling secure and scalable AI deployment.

4. High-Level Architecture

The architecture of the proposed AG-FAC is designed in an elegant way that it offers scalability and security to the machine learning deployments. The feature-level protections in the proposed architecture involve the use of AI to incorporate intelligence into each step of the model access lifecycle ranging from request intake, inference, and governance.

Highly at a glance, the architecture comprises six major building blocks including: (i) Request Interface Layer, (ii) Context & Identity Analyzer, (iii) Feature Sensitivity Registry, (iv) AI-Driven Risk Engine, (v) Policy Enforcement Layer, and (vi) Governance & Audit Module. The components form a loop that monitors and controls the entire process to ensure the continuous provision of secure feature access control.

The Request Interface Layer is responsible for the processing of requests for model inferences. Users, applications or external services submit requests using APIs. The Request Interface Layer validates the requests and sends them together with meta data to other components of the AG-FAC architecture. Meta data include user identity, request rate, and contextual factors.

The Context & Identity Analyzer analyzes behavioral and environmental indicators linked with each access request. The trust factors considered include user role, past behavior, location, device information, and access time. The analysis helps identify legitimate and suspicious behaviors based on contextual insights and serves as the foundation of access decisions.

The Feature Sensitivity Registry is an authoritative directory of features classified by their exposure risks, regulatory sensitivity, and business significance. The classification process uses explainability measures, compliance requirements, and corporate policies. The registry allows high-exposure features to be tightly controlled, whereas low-exposure features should not be blocked to maintain model effectiveness.

The AI-Driven Risk Engine combines contextual factors, feature sensitivities, and past accesses to generate an exposure risk score dynamically. The risk engine relies on advanced algorithms and machine learning to discover patterns and predict possible misuse cases, including model exfiltration, inference attacks, and data breaches. The results produced by this engine have a significant impact on enforcement decisions.

The Policy Enforcement Layer converts the calculated risk scores to tangible action by implementing the access decision according to either predefined or dynamically evolving security policies. Depending on the risk score, the enforcement layer can approve the request to access all features, limit access to certain features, obfuscate or add noise to some features, or even deny the entire request. Using the feature level rather than the model level helps achieve granular security without having multiple versions of the model.

Finally, the Governance & Audit Module offers continual oversight, documenting all access decisions, policy evaluations, and risk assessments. The module serves auditing and performance monitoring purposes as well as policy development. Feedback gathered from this layer is fed back into the risk engine and policy framework for continuous learning and improvement.

In summary, this overall architecture facilitates a secure, dynamic, and transparent process for controlling model exposure. Incorporating AI governance into the inference process streamlines an effective balance between security, performance, and business objectives in the enterprise setting.

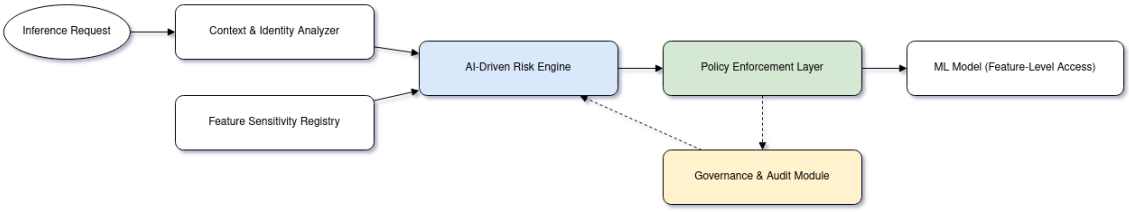

Figure 1. High-level architecture of the AI-Governed Feature Access Control framework, illustrating risk-aware feature enforcement and continuous governance across the model inference pipeline.

5. Market Opportunity

The growing deployment of AI in various industries like finance, healthcare, e-commerce, telecommunications, and governmental services has exponentially elevated the value of AI and the significance of its internal workings. With the increasing use of API access, SaaS offerings, and integrated decision-making processes in AI deployment, the exposure of AI model internal structure, feature leakage, and abuse becomes an important consideration for businesses. This creates a lucrative space for the development of AI protections that allow organizations to leverage their models without compromising scalability and efficiency.

The adoption of AI technology is quickly expanding, and more companies are now using their models in a variety of ways including multi-tenant clouds, cloud-native computing environments, and third-party systems. The access to models in these settings involves a broad range of users, each with varying degrees of trust. Conventional methods of access control are typically at the application or endpoint layer, which makes protecting the model internal components impossible.

Moreover, regulatory pressure enhances the need for such an innovation. In this case, frameworks like GDPR, HIPAA, ISO/IEC 27001, and even AI-related regulations emphasize the importance of granular controls, auditable process flows, and risk-based access to sensitive information and automated systems. Feature-level leakage represents compliance risks that are not mitigated through current MLOps practices or IAM tools. In turn, enterprises must rely on AI-native methods of governance that can change access decisions dynamically depending on the circumstances, risk profile, and behavior of users.

From a business perspective, AI governance can unlock new revenue models and use cases. Enterprises have the option of exposing different capabilities of models without having to duplicate them. As such, businesses can safely enable tiered AI services, facilitate secure model sharing, and collaborate with other organizations without any security risks. AI platforms and startups can especially take advantage of this innovation in order to protect their intellectual property and scale up their products.

In new digital environments that are developing rapidly in terms of cloud services, but which may exhibit varying degrees of security maturity levels, there will be an increased demand for automatic, intelligent, and efficient models protection measures. Feature access control with the help of artificial intelligence will fill this void through incorporating a governance mechanism into the prediction process.

On a more general note, the combination of AI proliferation, higher risks associated with exposure of AI models, regulatory pressure, and demands for revenue generation creates a compelling case for the proposed solution to emerge as a commercial venture.

6. Conclusions

The research was carried out with a view to addressing the increasing threat posed by model exposure risks in current state-of-the-art AI applications. The study proposes an AI-controlled feature access control mechanism that is capable of imposing fine-grained and context-aware restrictions on model feature utilization when running inference tasks.

Given that machine learning models are used more and more widely in distributed systems where there is not only a risk of exposure to malicious agents but also potential issues related to data ownership, traditional means of securing systems at the application/API-level become obsolete.

In order to solve these problems, the suggested feature access control policy allows organizations to dynamically adjust their security policies based on the current threat landscape, the users' trustworthiness, and their level of compliance with relevant standards. By applying this technology, one can effectively avoid the risk of unauthorized use of the model features, reverse engineering, and other malicious attacks against models and ensure intellectual property protection.

This research also points out the significance of feature-level governance from an organizational and regulatory standpoint. Selective exposure/restriction of model capabilities enables secure AI monetization, controlled collaboration, and compliance with data privacy and AI governance regulations that are yet to be fully implemented. This is especially relevant in industries processing sensitive or regulated data and countries that adopt AI quickly but lack advanced cybersecurity measures.
Although the proposed framework proves itself conceptually and architecturally feasible, further work will focus on empirical evaluation in realistic deployments, measuring its security impact quantitatively, and benchmarking its performance. Other possible areas of study include incorporating explainability-aware access policies into the framework, applying it in federated and edge AI settings, and assessing compatibility with current MLOps and IAM solutions.

In summary, AI-powered feature access control emerges as a vital step towards secure and scalable AI deployment. By moving security enforcement from coarse-grained endpoints to smart features, the proposed framework provides a realistic and forward-thinking tool for minimizing exposure risks in next-generation AI systems.

References

Amanna, A., Deploying next-generation artificial intelligence ecosystems for real-time biosurveillance, precision health analytics and dynamic intervention planning in life science research, Magna Scientia Advanced Biology and Pharmacy, vol. 16, no. 1, pp. 38–54, 2025.

Bakinde, A., and Ajibade, O., Bridging finance and operations: Automating cross-functional insights with SQL, ETL, and visualization tools, 2025.

Blašković, L., et al., Robust clinical querying with local LLMs: Lexical challenges in NL2SQL and retrieval-augmented QA on EHRs, Big Data and Cognitive Computing, vol. 9, no. 10, 256, 2025.

Dave, E., et al., Federated learning and MLOps pipelines: Driving privacy-preserving AI deployment at scale, 2025.

Dave, E., Adeola, F., and Dave, N., Advancing trustworthy AI in the cloud era: From generative models to privacy-preserving MLOps, 2025.

Esther, F. A., Kingsley, A., and Huston, F., Federated learning and MLOps pipelines: Scaling privacy-preserving AI in enterprise environments, 2025.

Harrer, S., et al., Artificial intelligence drives the digital transformation of pharma, in Artificial Intelligence in Clinical Practice, pp. 345–372, Academic Press, 2024.

Jain, S., et al., Need of anthropogenic intervention in achieving sustainable development (IC-NAIASD-2023), 2023.

Musunuri, A., Developing a scalable AI framework for moderating social media content, International Journal of Computer Applications, 2025.

Noel, D., Adeola, F., and Jacobs, I., Advancing enterprise AI: From generative models to privacy-preserving systems and secure MLOps pipelines, 2025.

Stark, J., and Gill, A., From funding to trust: Australia's sovereign AI architecture, Authorea Preprints, 2025.

Zulfikaroglu, B., and Ozmen, M. M., Artificial intelligence in surgery: From anatomy to ethics, 2025.