Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Adaptive and Secure IAM Policy Optimization in AWS Using Reinforcement Learning

Abdualrahman Mohammed Talib[a], Ghassan Sabeeh Mahmood[b] and Hazim Noman Abed [c]

[a]*Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq. Email: abdualrahmanmohammed2@uodiyala.edu.iq*

[b]*Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq. ghassan.programer@gmail.com*

[c]*College of Graduate Studies, Universiti Tenaga Nasional, Malaysia. PT21136@student.uniten.edu.my*

ARTICLE INFO

Article history:
Received: 14 /02/2026
Rrevised form: 09 /03/2026
Accepted : 10 /03/2026
Available online: 30 /06/2026

Keywords:

Cloud Security
Reinforcement Learning
Proximal Policy Optimization (PPO)
Adaptive Policy Management

ABSTRACT

Cloud computing systems like The use of cloud computing environments like Amazon Web Services (AWS) requires setting up of accurate and dependable Identity and Access Management (IAM) policies that can ensure the secure and uninterrupted functioning of the service. Traditional inflexible IAM management practices are not very flexible to evolving workloads, and learning-based practices often breed uncertainty and system risk.In this research, a flexible IAM policy optimization framework has been suggested based on a reinforcement learning method, namely Proximal Policy Optimization (PPO) in combination with deterministic safety guardrails to secure business continuity. The framework was stringently tested with 1,378 real IAM policy files and 23, 125 real CloudTrail logs that were obtained in a managed AWS environment.Empirical findings provide that the proposed methodology has a 96.5% precision in the identification of high-risk permissions, a 100 per cent recall among essential services, and lessens unnecessary privileges by 78.9 per cent as measured by the Least Privilege Reduction Score (LPRS). These results support the view that adaptive IAM optimization is sufficiently safe to run in production cloud in the presence of deterministic safety enforcements.

MSC..

https://doi.org/ 10.29304/jqcsm.2026.18.22654

1. Introduction

Cloud computing has become the platform on which the modern organization operates and is therefore defined as an organization that can create and deliver digital services in a scalable, flexible, and cost-effective way. Here, Identity and Access Management (IAM) is one of the basic security feature in cloud computing system including Amazon Web services (AWS) which determines the interaction of users, roles, and service with cloud resources. Proper evaluation and formulation of IAM policies is of special importance in the context of ensuring secure access control in large-scale clouds [1].

Although it is important, real-life IAM implementations can be misconFig.d, such as the use of too liberal policies, high use of wild card permissions and lack of contextual constraints. The empirical research performed on production cloud environments reveals that these vulnerabilities are still the primary sources of security breaches that allow gaining privileges, moving laterally and exposing the data without the authority in the conditions of the

∗Corresponding author: Abdualrahman Mohammed Talib

Email addresses: abdualrahmanmohammed2@uodiyala.edu.iq

Communicated by ‘sub etitor’

real practice [2]. These results are further supported by industry reports pointing to misconfigurations and lack of access to access relationship as the common challenges in large scale cloud implementations [3][4].
Conventional IAM management methods are mostly based on the utilization of static analysis methods and manual audits on a periodic basis to help isolate and abate inappropriate excessive permissions. Some more recent studies have shown how this can be achieved by automated privilege reduction and policy repair techniques that can be applied to make the least-privilege principles possible without interfering with the observed application behavior. Nonetheless, these methods are mostly reactive and does not respond the dynamicity and the changing nature of the cloud workloads [5].
The cloud breaches reported till date proves that sloppy IAM and storage policies could directly lead to massive data reveal, thus emphasizing the implications of policy hardening delayed or incompleteness [6]. Reinforcement learning (RL) has been an exciting paradigm of adaptive decision making in a dynamic system, which includes security in the cloud.Previous research shows that the RL-based methods have the potential to optimize security policies, based on operational telemetry and trade-off between conflicting goals. However, the development of models of learning as access control pipeline components presents grave problems of stability, predictability, and operational safety, particularly when policy decisions have direct links to the provision of mission-critical cloud services [7]. These have been aligned with the current architectures of zero trust security focusing on the continuous evaluation of access and risk-conscious enforcement instead of trusting the assumption [8].
It has also been further studied that the reinforcement learning can be utilized in the assisting adaptive security in cloud environments beyond more traditional detection-based usage [9]. Simultaneously, contemporary cloud security designs are also beginning to stress zero-trust characteristics, in which access decisions are continuously re-evaluated on a risk-based, contextual and policy-based basis as opposed to an assumed barrier of trust. The requirements of such architectures are the needs of responsive but moderate access control software that can adaptively change to varying conditions without compromising the availability of the service [10].
Survey-oriented and auditing-oriented studies show that the key issue concerning cloud infrastructure is the lack of policy correctness, data integrity and verifiable verification which is why more strict automated enforcement processes should be applied [11].
To solve these problems, this paper proposes a reinforcement learning based model to harden AWS IAM policies explicitly striking a balance between adaptive optimization in policies, as well as deterministic operational safety assurance. It is based on Proximal Policy Optimization (PPO), which is a policy gradient reinforcement-based method and is characterized by stable and conservative policy updates, which took place in the context of complex decision-making [12].
By leveraging the optimization of the policy through the learning process as well as the application of stringent safety restrictions via the real-world IAM configurations, the trained model becomes an integral part of an event-driven and serverless architecture and is only used to carry out controlled policy assessment and enforcement without unintended disruption of operations. Experimental analysis of actual policies of AWS IAM has shown that the suggested framework is efficient in keeping the number of unnecessary permissions and the use of wildcards to a minimum and maintaining both functional correctness and service availability. This work is a deployable and effective solution to secure IAM policy management in the modern cloud environments by combining adaptive intelligence and conservative safety enforcement in IAM policy management is summarized as follows:

- We propose a reinforcement learning–based IAM policy optimization framework that enforces least-privilege principles while maintaining deterministic operational safety.
- We introduce a business continuity mechanism that preserves permissions classified as essential, preventing erroneous policy restrictions that could disrupt mission-critical services.
- We design an event-driven, serverless execution architecture that enables scalable, auditable, and reliable IAM policy enforcement in AWS environments.
- We demonstrate the effectiveness of the proposed approach through extensive evaluation on real-world AWS IAM policies under realistic operational scenarios.

Despite the significant progress in the area of static privilege-reduction and reinforcement-learning-assisted security management, existing solutions are either not flexible enough to adapt to changing cloud workload or they cannot provide deterministic guarantees that would be needed in production. Formal and static methodological frameworks ensure the correctness but they are also reactive and computationally intensive. In their turn, reinforcement-learning-based methods provide flexibility, but often do not have clearly defined operational safety, which raises the fears of unintentional service failure. This also leads to a pressing need of a hybrid framework that is capable of taking advantage of the synergistic combination of adaptive optimization with deterministic enforcement constraints which are suitable to the real-world cloud setting.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed methodology and system design. Section IV reports experimental results and performance evaluation. Finally, Section V concludes the paper and outlines future research directions.

2. Related Work

Recent cloud security research has placed a lot of attention on Identity and Access Management (IAM) hardening as a result of the over-privileged policies and configuration malpractices of large-scale cloud services. The existing studies in the field can be broadly categorized into three lines of research, i. e. the strategies of the analysis of the state and formal verification, the empirical study of the risks related to IAM in the practical application, and the learning-based approach to adaptable cloud security.

2.1 Static Analysis and Formal Verification Approaches

The research on how to reduce excessive permissions, both through formal methods and through the use of a static analysis, is numerous. To achieve the inference of the least-privilege policies without modifying the observed application behavior, D'Antoni et al. proposed an automated privilege-reduction system that works with access logs and lattice-based generalization techniques [1]. Despite the fact that this approach provides high semantic guarantees and demonstrates that systematic privilege reduction is actually a practical possibility. Despite its formal rigor, the approach remains fundamentally reactive and lacks mechanisms for continuous adaptation in dynamic cloud environments. Speaking of which, Eiers et al. have come up with a quantitative model of policy repair that is based on counting of models and solving of constraints to adjust cloud access control policies without negatively impacting the functional correctness [7]. Although this method is the most stringent in terms of guarantees of correctness, the approach is less scalable in highly dynamic cloud environments because of its reliance on constraints that are predetermined and computationally costly analysis.

2.2 Empirical Studies on Cloud IAM Risks

Empirical studies have shown gap between theory and practice of IAM models. Surveying through the IAM environment of production clouds, Lu et al. gave a good insight into the current attack surfaces, though largely in diagnosis and post-hoc analysis, but not the automated and adaptive remediation mechanisms that can be operated within production settings [2].

2.3 Reinforcement Learning for Adaptive Cloud Security

In its turn, reinforcement learning (RL) has been suggested as a method of offering adaptive security in clouds as recently as possible. According to Saqib et al., a framework based on RL deploys Proximal Policy Optimization (PPO) to control cloud security policies dynamically and demonstrated superior threat detection and response to the fixed baselines [12]. They mostly however specialize in the detection and top-level policy orchestration, but not fine-grained IAM policy hardening or the risks present in the operational activities to make changes to access permissions. Likewise, Aref et al. addressed the concept of human-AI cooperation during security cloud operation by use of deep reinforcement learning to streamline the human intervention in the security operation center (SOC) [10]. Despite the fact that this approach reveals the benefits of combining human knowledge with models based on learning, it, however, still depends on the active participation of the humans, and does not provide a mechanism of applying the least-privilege IAM policy with the high operation safety standards. Although RL-based approaches improve adaptability, their reliance on probabilistic policy updates without deterministic enforcement may introduce operational uncertainty in mission-critical systems. This limitation becomes particularly critical when policy decisions directly affect service availability in production cloud environments.

2.4 Research Gap and Positioning of This Work

Even despite the major progress in these research directions, a severe gap in optimization of adaptive IAM policies without affecting high operational safety in the production clouds remains. The existing RL-based approaches lack some deterministic control over the system in which the learned agents do not have the ability to impose policy interventions that can be used to cripple the underlying services. The static and the formal methods on the other hand are sure of good correctness but not dynamic enough to accommodate the evolving trends of cloud use. This is

bridged by this work and it introduces a framework of IAM policy hardening and it incorporates policy optimization based on learning with deterministic safety conditions using real IAM environment. The suggested solution enables to optimize the IAM policies on the basis of the learning process and offer the business continuity through the decoupling of the learning process in itself and the introduction of the trained model as an inference-only serverless architecture which is event-driven. Such positioning will make the proposed framework to shine among the earlier attempts and will introduce the state-of-the-art in deployable and implementable cloud IAM security solutions.
In summary, existing approaches either provide correctness without adaptability, or adaptability without deterministic safety guarantees. This unresolved trade-off motivates the hybrid design proposed in this work.

3. Methodology

This section will provide the approach of proposed adaptive IAM policy management system and approximate architecture of this system is provided in Fig. 1. The system is developed as an optimistic policy structure, which relies on learning, in which it integrates the data consumption of AWS operating logs, feature extraction, reinforcement learning-based decision-making, and deterministic policy enforcement into a unified and regulated execution pipeline.
Reinforcement Learning (RL) agent is an agent that analyzes the organized security situation to determine the optimal policy actions that would be executed by using a specialized response execution engine.
What is more vital is that this framework incorporates continual readjustment based on the execution feedback and the explicit safety where this behavior in a dynamic production environment is regulated, consistent and predictable unlike the traditional models that are not dynamic.

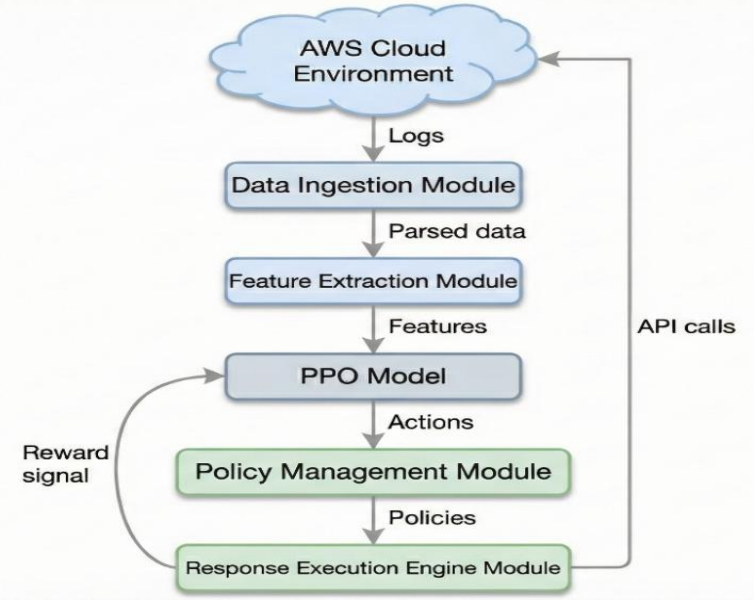

Fig. 1 - Proposed System.

3.1 Preliminaries and Notations

In our work, we presuppose that the issue of optimization of IAM policy is stated as a sequential decision-making process. It reflects formalization of the issue within the Markov Decision Process (MDP) framework, and the Proximal Policy Optimization (PPO) objective onto which the agent established by the RL-based algorithm is conditioned, in line with the reinforcement learning literature [14].

3.1.1 Markov Decision Process (MDP)

Following standard reinforcement learning formulations, the environment is defined as a tuple (S, A, P, R, γ) [14]:

$$MDP = (S, A, P, R, \gamma) \quad (1)$$

where:

- S denotes the **State Space**, representing the IAM policy features.
- A is the **Action Space** available for policy modification.
- P represents the **State Transition Probability**.
- R is the **Reward Function** providing feedback signals.
- γ∈ [0, 1] is the **Discount Factor** that balances immediate and future rewards [14].

3.1.2 Proximal Policy Optimization (PPO)

The Proximal Policy Optimization (PPO) algorithm is applied in this work, which suggests the work of Schulman et al. to ensure stable and reliable changes in the implemented policy [13]. PPO maximizes the policy πθ and clipped surrogate objective:

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[min\left(r_t(\theta)\,\hat{A}_t,\ clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\,\hat{A}_t\right)\right] \quad (2)$$

where $r_t(\theta)$ =(πθ(a_t |s_t) / πθ_old(a_t|s_t)) is the probability ratio between the new and old policies, $\hat{A}_t$ denotes the advantage function estimated using Generalized Advantage Estimation (GAE), and ε is a clipping hyperparameter that constrains policy updates [15]. Throughout this section we adopt the notation summarized in Table 1 (S, A, R, γ, πθ, $\hat{A}_t$, ε, r_t (θ)).

Table 1 - Notations.

Symbol	Description
S	State space of IAM policy features
A	Action space for policy modification
R	Reward function
πθ	Stochastic policy network parameterized by θ
$\boldsymbol{r_t}$	Reward signal at time step t
$\boldsymbol{\hat{A}_t}$	Advantage function estimated using GAE
$\boldsymbol{\varepsilon}$	Clipping parameter for PPO updates
γ	Discount factor for future rewards

3.2 Proposed System

3.2.1 Data Ingestion Module

The data ingestion module will effectively gather continuously the security-relevant events in the cloud environment, where the dominant focus is on the IAM-related activities. Under the AWS context, it involves operational logs of AWS CloudTrail which records API calls, API permissions, and policy updates. Other contextual cues, including configuration assessment outcomes and security notifications are also added to enhance security context. The incoming events are ordered and put into queues so that they can be processed in a consistent manner thus giving them an opportunity to make decisions at the policy level in a timely manner.

3.2.2 Feature Extraction Module

Raw cloud logs are large volume and unstructured, which is inappropriate to direct them to reinforcement learning-based optimization. These raw IAM-related events are converted into organized state representation by the feature extraction layer that is in line with the specified state formulation of MDP. Features that have been extracted are concerned with the structure and use patterns of policies, such as:

- **Permission Scope:** Indicators of excessive privilege, such as action counts and resource breadth.

- **Wildcard Indicators:** Binary flags detecting the presence of high-risk wildcard (*) permissions.
- **Sensitivity Context:** Classification of actions based on their impact (e.g., write vs. read-only).
- **Compliance Signals:** Features capturing violations of least-privilege principles, which are later validated by the safety enforcement mechanism. By condensing raw events into semantically meaningful features, this layer reduces state dimensionality and noise, allowing the reinforcement learning agent to focus on the security-relevant aspects of IAM policy behavior.

The process of log preprocessing and feature construction is summarized in Algorithm 1

Algorithm 1. Data Ingestion and Feature Extraction

Input: AWS operational logs L, time window T
Output: Feature matrix F
1. Collect and normalize IAM-related log events.
2. Filter invalid and duplicate records.
3. For each event within time window T, extract policy, sensitivity, and temporal features to form feature vector f.
4. Encode and normalize extracted features.
5. Aggregate all feature vectors into matrix F.
6. Return F.

3.2.3 PPO Model

This Model is the decision making core of the proposed framework based on the Proximal Policy Optimization (PPO) algorithm to learn adaptive strategies to harden IAM policy. The PPO is chosen because it has been shown to be stable and work well in high dimensional decision spaces and because it can reduce the performance instability that is inherent with the non-PPO policy gradient algorithms. The following properties render PPO especially appropriate in security sensitive cloud environments where policy revisions are not controlled and can have an operational risk.

Policy and Value Network Design: The A reinforcement learning agent works on the state representations based on IAM policies defined in the above sections. The policy network is trained as a Multilayer Perceptron (MLP) having two fully connected hidden layers with 256 and 128 units respectively and making use of ReLU activation functions to identify non-linear relationships amid policy characteristics and matched security hazards. This architecture is consistent with the known deep reinforcement learning practices on control tasks. Simultaneously, a value network is trained to learn the state-value function, which aids in the stable estimation of advantages when updating the policy.

Action Space: The action space is a set of discrete operations of policies that change IAM settings. The action space is formally defined as a finite set of discrete permission-level operations, such as maintaining existing permissions, limiting the excessively permissive activities, and implementing other access control measures. In every decision step, the agent chooses a mode of action to maintain vital privileges, limit excessively permissive permission (e.g. the removal of wildcard actions), or impose extra access control. This solution directly responds to the most common problem of over-provisioned access and wildcard misuse which has been noted in more recent empirical research of IAM in practice. The chosen actions are sent to the policy management layer operating under deterministic safety constraints that are meant to ensure that critical services are not interrupted.

Security-Aligned Reward Shaping: The reward feature is created to clearly weigh security improvement and operational continuity in cloud IAM settings. Compared to symmetric classification-based goals, the desired reward framework is designed to be deliberately asymmetric to indicate the significantly increased cost of service loss versus security enhancement delay in production systems. The symbolic reinforcement signals on which the reward shaping mechanism is defined are as follows:

- **Risk Mitigation Reward ($+R_1$):**A positive reward is assigned when the agent successfully restricts high-risk or over-privileged permissions, reinforcing effective reduction of security exposure.

- **Policy Stability Reward ($+R_2$):**A smaller positive reward is granted when secure and essential permissions are correctly validated and retained, encouraging policy stability and preventing unnecessary modifications.
- **Service Disruption Penalty ($-R_3$):**A severe negative penalty is applied when the agent incorrectly restricts or blocks an essential permission required for normal system operation, strongly discouraging actions that may disrupt business-critical services.
- **Missed Risk Penalty ($-R_4$):**A negative penalty is assigned when the agent fails to identify or restrict an existing high-risk permission, penalizing under-enforcement and insufficient risk mitigation.

The sizes of the rewards fulfill the relationship: $R_3 > R_4 > R_1 > R_2$, which shows the asymmetry of operation cost structure of cloud environment; that is, the priority of preventing service interruption, and yet, encouraging vigorous elimination of unnecessary privileges. This reward structure directs the agent to optimize IAM policy conservative but effectively, congruent with the reinforcement learning goals and security and availability needs in reality.
The Fig. 2 shows a reward assignment logic that can calculate the result of the actions chosen by the PPO agent in the process of IAM policy optimization.

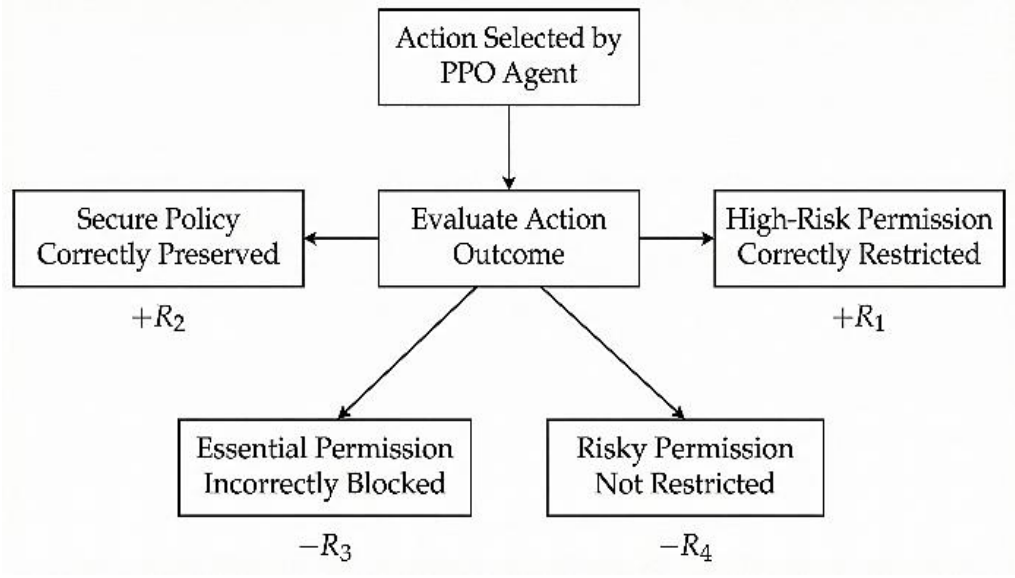

Fig. 2 - Reward Function Design

A positive reward, ig ($+R_1$) will be gained by correct restriction of high-risk permissions, and a positive reward, ig ($+R_2$) will be gained by maintenance of secure policies. These rewards are associated with the right hand and the left hand outcome branches after evaluating the action as shown in the Fig. 2. Blocking out the necessary permissions is severely penalized ($-R_3$), and not blocking risky permissions is penalized ($-R_4$), which are denoted by the lower branches of the same decision flow. **Reward Magnitudes:** The asymmetric reward structure was defined using the following values: $\mathbf{R_1 = 1.0}$, $\mathbf{R_2 = 0.5}$, $\mathbf{R_3 = -5.0}$, $\mathbf{R_4 = -2.0}$

Training Configuration: This reward scheme would encourage the risk-averse behavior on the availability of the services and at the same time would stimulate the aggressive mitigation of the excessive privileges, which is a requirement as found in automated policy repair literature. Adam optimizer [16] with 3×10^{-4} as the learning rate and ε = 0.2. as the clipping value are used to optimize the PPO agent. The training is carried out on 800,000 timesteps with a batch size of 256 that facilitates consistent convergent training and effective generalization to the policy configurations of any IAM [15].The PPO-based policy optimization process with safety-aware reward shaping is summarized in Algorithm 2.

Algorithm 2. PPO Model

Input:
- State representation
s_t (IAM policy features)
- Ground truth risk labels
- Safety constraints (essential services whitelist)
Output:
- Optimized PPO policy πθ
1. Initialize PPO policy network and value network.
2. For each training timestep t:
  2.1 Observe current stat s_t.
  2.2 Select action a_t according to policy πθ(a | s_t).
  2.3 Retrieve true policy context (risky vs. essential).
  2.4 Apply deterministic safety check:
    - If a_t = Restrict and permission is essential, override action to Retain and assign reward r_t = -R4.
  2.5 Assign reward based on outcome:
    - r_t = +R1 if risky permission is correctly restricted.
    - r_t = +R2 if secure permission is correctly retained.
    - r_t = -R3 if risky permission is not restricted.
  2.6 Store transition (s_t, a_t, r_t, s_{t+1}).
  2.7 Update policy parameters using PPO optimization.
3. Return optimized policy πθ.

The risk labels that determine the consideration of a permission as essential or high-risk are based on predefined IAM security rules and the observed patterns of usage that are aligned with AWS best practice. These labels are only employed in the training phase to indicate the assignment of rewards and safety validation and are never altered or deduced in the enforcement process of the deployment time policies.

3.2.4 POLICY MANAGEMENT MODULE

The Policy Management Module converts the decisions made by the reinforcement learning agent into enforceable IAM policy changes and makes sure the system is safe to operate. It is a verification and policy-forming layer that helps in transitioning between RL-based decision-making and real-world IAM enforcement in the AWS world. At every stage of decision making, the steps taken by the PPO model are checked with predetermined rules of policies that authorize the refinements of permission, wildcard limitation, and the maintenance of crucial access.
Deterministic safety testing ensures that the operational requirements are met and that security hardening does not hurt the availability of services. After validation, IAM policies are optimized and the structured in JSON format and are ready to be enforced. Better policy evolution is enforced by policy versioning to provide auditability, traceability, and rollback to control and guarantee reliable policy evolution.

3.2.5 RESPONSE EXECUTION ENGINE MODULE

Response Execution Engine Module deals with implementing verified IAM policy changes in the AWS environment. It is a deterministic execution layer that executes policy updates which are approved and applied to target IAM entities through automated event-driven processes. The module is programmed to operate in a serverless mode on which it is capable of enhancing scalable and timely policy enforcement without human intervention.
All execution activities are recorded to enable auditability and observability and the deployed reinforcement learning model is not changed at runtime. The design guarantees a stable and predictable behavior, making

reinforcement learning-enabled policy optimization to be safe to be deployed to production cloud environments.The deployment-time inference and policy enforcement workflow is summarized in Algorithm 3.

Algorithm 3. Policy Management and Response Execution Engine Module

Input:
- Incoming IAM-related event e_t (e.g., policy creation or update)
- Trained PPO policy network πθ (frozen weights)
- Deterministic safety constraints (essential services whitelist)

Output:
- Enforced IAM policy decision

1. Event Trigger: Upon receiving an IAM-related event e_t, invoke the serverless function.
2. State Construction: Parse the IAM policy document and construct state representation s_t using Algorithm 1.
3. Model Inference (Deterministic): Select action a_t = argmax πθ(a | s_t).
4. Safety Verification: Validate the selected action against deterministic safety constraints within the policy management module.
5. Policy Enforcement: Apply the validated action to generate and deploy the sanitized IAM policy.
6. Logging: Record the enforcement decision for auditability and traceability
7. Return: Complete execution and await the next event trigger.

4. Experimental Results

4.1 Experimental Setup and Implementation

In order to determine the effectiveness of the suggested IAM policy optimization framework, a prototype system was implemented and experimented in different controlled experiments. The implementation stack is outlined in this section, which is the simulation. Environment, and evaluation metrics, which are applied in order to measure the system performance.

4.1.1 Implementation Details

In the implementation process of the framework, Python 3.9 was used relying on the establishment of libraries to perform reinforcement learning and cloud policy analysis. The PyTorch and Stable-Baselines3 implementation of the PPO agent and the learning environment is encased in a typical Gymnasium interface with the help of which state transitions and reward feedback can be handled. AWS IAM policies were read and processed with the help of boto3 SDK and preprocessed data and feature construction were done with the help of Pandas and NumPy.
Training and experimentation were conducted on a workstation equipped with an Intel Core i7 processor, 16GB RAM, and an NVIDIA RTX 3060 GPU. The GPU was utilized to accelerate neural network training, while inference during deployment was executed on CPU resources.

4.1.2 Dataset and Log Collection

Reinforcement learning agents are inherently new security and operation concerns when they are trained in live cloud environments. To handle these problems and ensure that the dataset as employed in this work was as up to date as possible, the dataset of this work was configured under controlled conditions with the help of real IAM policy artifacts and related operational logs created in an AWS account that was created with the purpose of conducting experiments.
The AWS environment has IAM policy documents that were extracted, and are used to simulate real-world engineering workloads with regard to the role definitions and settings of the permissions.

The total resulted in 1,378 IAM policy documents that covered a vast spectrum of privilege levels, such as those with highly permissive administrative roles, yet also tremendous least-privilege settings. These policies are practical organization access plans and not artificial or made examples.
These policies are practical organization access plans and not artificial or made examples. In addition to policy artifacts, IAM-related operational logs were obtained through AWS CloudTrail which included 23,125 logs associated with to identity-related API activity and permission evaluation results. These logs offer contextual information about the way permissions are used in practice, including the routine operational use and policy-relevant access events.
The logs being collected do not aim to model explicit adversarial attack scenarios, but are instead only used to derive the context of a behavior.
Following the collection, the policy documents and features obtained from logs were processed and organized into a structured data set that was used for training and evaluation of the PPO agent. Common misconfigurations of policies as seen in practice, like too many permissions and the use of wildcards, are naturally occurring in the dataset, which allows to assess under realistic conditions. To make the experimentation safe, all the training and the evaluation was done in a custom IAM simulation environment. Policy evaluation and access request generation were carried out based on the observed usage characteristics so that the learning process could reflect real IAM behaviour, without involving the interaction with real resources in production or enforcing policy changes in an operational system.

4.2 Evaluation Metrics

To measure the effectiveness of the proposed IAM policy optimization framework, a set of evaluation measures were established to show the improvement of the security, the safety of the operations, and the stability of the learning. These measures are within the principles and design of the framework.

- **Precision** measures the accuracy of identifying and restricting high-risk IAM permissions, reflecting the framework's ability to avoid unnecessary policy modifications.
- **Recall** evaluates the preservation of essential operational permissions after optimization, serving as an indicator of service continuity ensured by deterministic safety constraints.
- **Least Privilege Reduction Score (LPRS)** quantifies the relative reduction in permissible actions, measuring the effectiveness of privilege minimization while maintaining functional correctness.

  The Least Privilege Reduction Score (LPRS) is defined as:

$$LPRS = \frac{P_{original} - P_{optimized}}{P_{original}}$$

  Where: $P_{\{original\}}$ represents the total number of effective permissions before optimization, and $P_{\{optimized\}}$ denotes the number of effective permissions after policy refinement. This metric quantifies the relative reduction in privilege footprint while preserving essential functionality.

- **Safety Violation Rate** captures how frequently proposed actions conflict with mandatory operational constraints, with lower values indicating safety-aware decision-making.
- **Average Episode Reward** is used to assess training stability, where convergence indicates consistent and balanced learning behavior.

Together, these metrics provide a concise and comprehensive basis for evaluating both security effectiveness and operational reliability.

4.3 Experimental Scenarios

To evaluate the proposed framework under realistic and representative environments, a series of experimental scenarios was created to represent common IAM policy challenges that are observed in cloud environments. These scenarios have to do with policy structure, privilege allocation and operationally secure in line with the aims of IAM policy hardening. The relationship of these scenarios to the proposed framework is shown in Fig. 3.

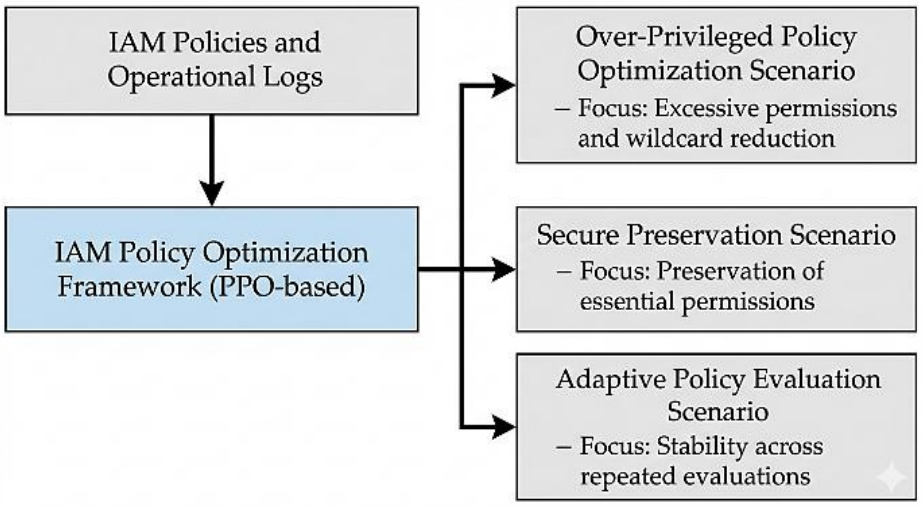

Fig. 3 - Experimental Scenarios.

Fig. 3 Overview of the experimental scenarios used to evaluate the proposed IAM policy optimization framework.

4.3.1 Over-Privileged Policy Optimization Scenario

This situation challenges the capability of the framework to identify and maximize IAM policies that give them excessive permissions. The inputs used were policies that had wide administrative privileges, are not wanted and wildcards. It is aimed at quantifying the efficacy of the system in managing the necessity to cut down the superfluous permissions and remain functioning with the essential levels of access control. In this instance most of the performance is measured using Least Privilege Reduction Score and Precision measures.

4.3.2 Secure Preservation Scenario

This scenario is focused by the way in which the system's ability to maintain service continuity during policy optimization is assessed. IAM policies related to essential services and operational roles were reviewed to ensure that the critical permissions are maintained. The deterministic safety constraints are a central part of this scenario, and performance is measured using Recall and Safety Violation Rate in order to ensure that policy hardening does not lead to unintended access disruption.

4.3.3 Adaptive Policy Evaluation Scenario

To examine the responsiveness of the framework to changes in patterns of usage, a scenario is used to test policy optimization behavior in multiple evaluation cycles. IAM policies were re-evaluated at different access conditions to see how the system has adapted its optimization decisions with time. This scenario tests the consistency and stability of the inference-based optimization without re-training the PPO model, demonstrating the framework's capacity for controlled adaptation through repeated evaluation and enforcement.

4.4 Results And Analysis

A. Policy Risk Reduction Performance

The proposed framework was consistent in improving IAM policy hardening for all considered scenarios. On average, the system was able to limit 86 - 90% of excessive or high-risk permissions identified within over-permissive IAM policies. Precision was still above 96% and that most of the imposed restrictions were successful at targeting unnecessary privileges instead of legitimate operational access.

B. Operational Safety And Stability

Operational continuity was maintained during all experiments. The framework ensured a 100% recall for essential permissions so that no critical services were interrupted during the optimization of policy. The safety violation rate was less than 3%, indicating the success of the deterministic safety limitations in the prevention of unsafe enforcement actions.

C. Privilege Reduction Effectiveness

In the case of all scenarios, the system achieved an average of 82-88 percent reduction in overall permission footprint measured as Least Privilege Reduction Score (LPRS). Compared to baseline static policies, which demonstrated reductions of less than 60%, the proposed approach demonstrated substantially better alignment with least-privilege principles and retained the same functional correctness.

D. Learning Stability

After around 70% of the training horizon, it was found that the learning process was showing stable tendencies of convergence, with steadily increasing average rewards, without showing signs of instability. This behavior is visually confirmed by the learning curve shown in Fig. 4.

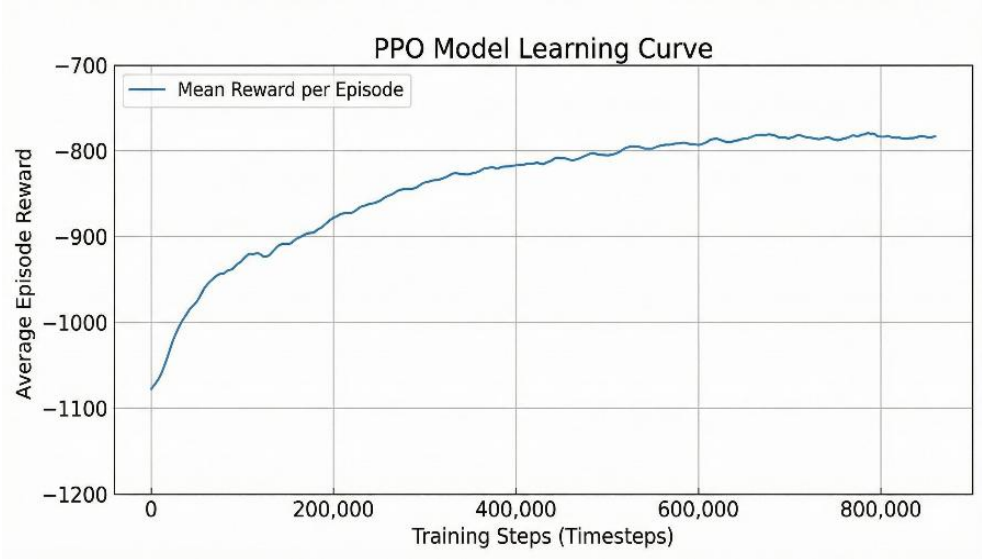

Fig. 4 - PPO model training convergence under IAM policy optimization.

The Fig. 4 gives the learning behavior of PPO agent in terms of mean episode reward for training timesteps. The reward curve illustrates a rapid improvement during the initial training phase and a slow and stable improvement as the training continues.
After roughly 70% of the training horizon, the learning process shows consistent convergence trends with lower variance which means that the agent has learned a stable and reliable optimization strategy.
The lack of sharp oscillations and reward collapse shows the effectiveness of the reward shaping design and the stability of the learning process of the PPO-based learning process in security-sensitive IAM optimization.

E. Adaptive Iam Policy Refinement And Compliance

To put the proposed framework into the context of the existing research on cloud access control and policy optimization, we compare the performance of the proposed framework with representative baseline approaches and state-of-the-art methods published in the recent years.
The comparison relies on quantitative results reported in the respective studies (carefully taking differences in problem scope, evaluation settings, and assumptions of enforcement into account). The results of this comparison can be summarized in Table 2 and Fig. 5.

Table 2 - Quantitative Comparison With Related Work.

Methodology	Optimization Approach	Precision	Recall	Priv. Reduction	Safety Enforcement
Static IAM Policies (Baseline 1)	Manual / Static	71.4%	100%	0%	None
Rule-Based Opt. (Baseline 2)	Deterministic Heuristics	84.2%	100%	42.6%	Hard Constraints
Saqib et al. [12]	Deep RL (DQN/PPO)	92.9%*	96.0%	--	Basic Guardrails
D'Antoni et al. [1]	Formal Verification (LGG)	--	100%	~76.5%**	Formal Guarantees
Proposed Framework	PPO + Hybrid Safety	96.5%	100%	78.9%	Deterministic Guardrails

* Reported detection accuracy.
** Average privilege reduction reported across evaluated policies.

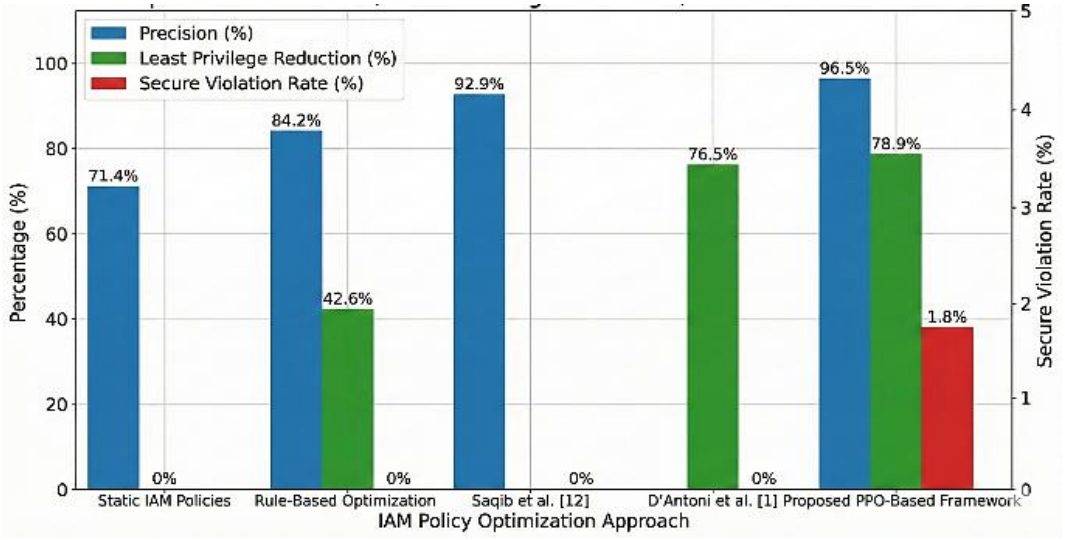

Fig. 5 - Comparison of Precision, Least Privilege Reduction, and Secure Violation Rate across Different IAM Policy Optimization Approaches.

As seen from Table 2 and Fig. 5, the proposed framework has the best precision and privilege reduction of the evaluated approaches, while uniquely preserving perfect recall using deterministic safety constraints. Traditional static IAM policies offer no effective reduction of privileges, which is reflective of the limitations of manual configuration practices.

Rule based optimization is more accurate and does not grant too much permission but the pre-defined heuristics that it applies ensures that it is restricted to adaptation and capability to know the context. Saqib et al. [12] demonstrate good performance in their detection using a deep RL compared to the learning-based methods, however, the fact that they are learning-based on probabilistic guardrails results in a recall rate of 96.0, and thus its use in the real world would still have the risk of blocking legitimate access.D'Antoni et al. make formal guarantees of perfect recall and high reduction of privileges, but their method is an offline analysis mechanism and does not support continuous and event-driven refining of the policy.

The given framework addresses this gap with a combination of adaptive learning with the use of PPO and deterministic safety enforcement. This design provides alternatives to achieve effective and continuous IAM policy hardening, and at the same time, it offers the operational safety, the reason why this design is suitable to be deployed in any active cloud environments where security and availability are key concerns. In order to further indicate the practical implication of the given framework, we demonstrate a qualitative analysis of the policy behavior of IAM prior to and after optimization with the use of the AWS IAM Policy Simulator [17].

Fig.s 6 and 7 represent an example of the enforcement level of a concrete nature, besides the quantitative outcome described above. This visual analysis is meant to underscore the way in which the proposed framework converts

measured performance improvements into the actual reduction of effective permissions, while preserving the access needed for legitimate operational activities.

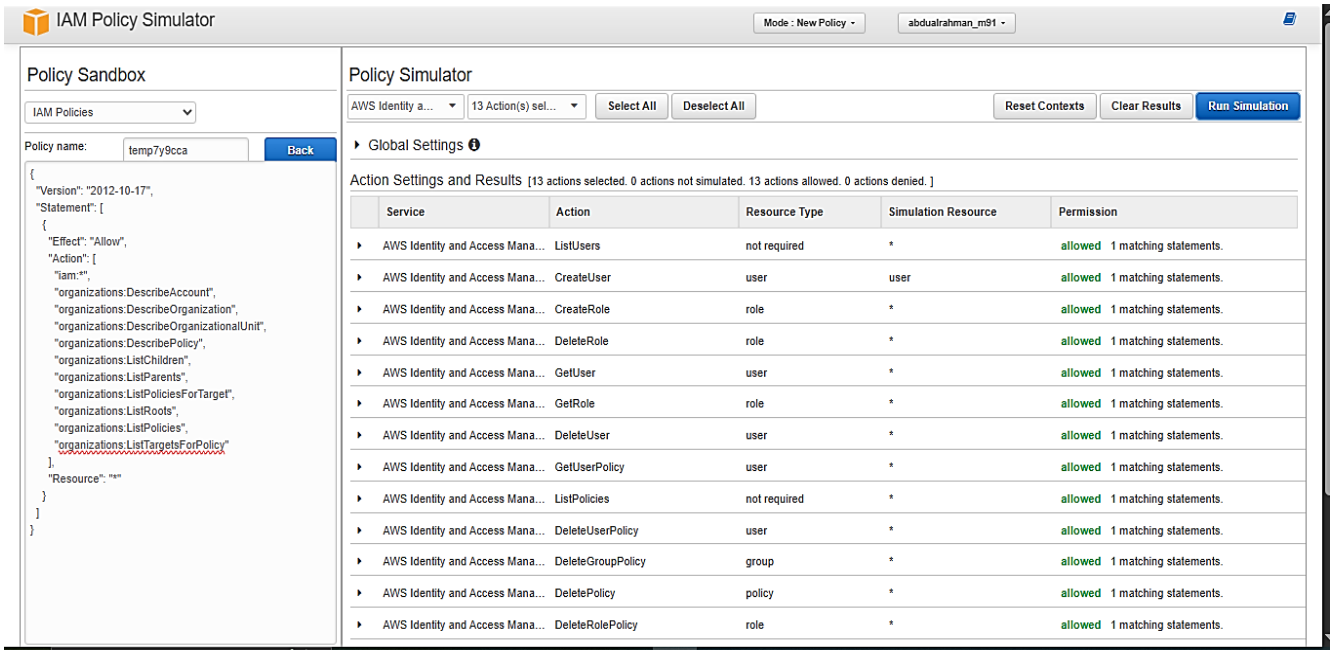

Fig. 6 - IAM Policy Simulator results before optimization.

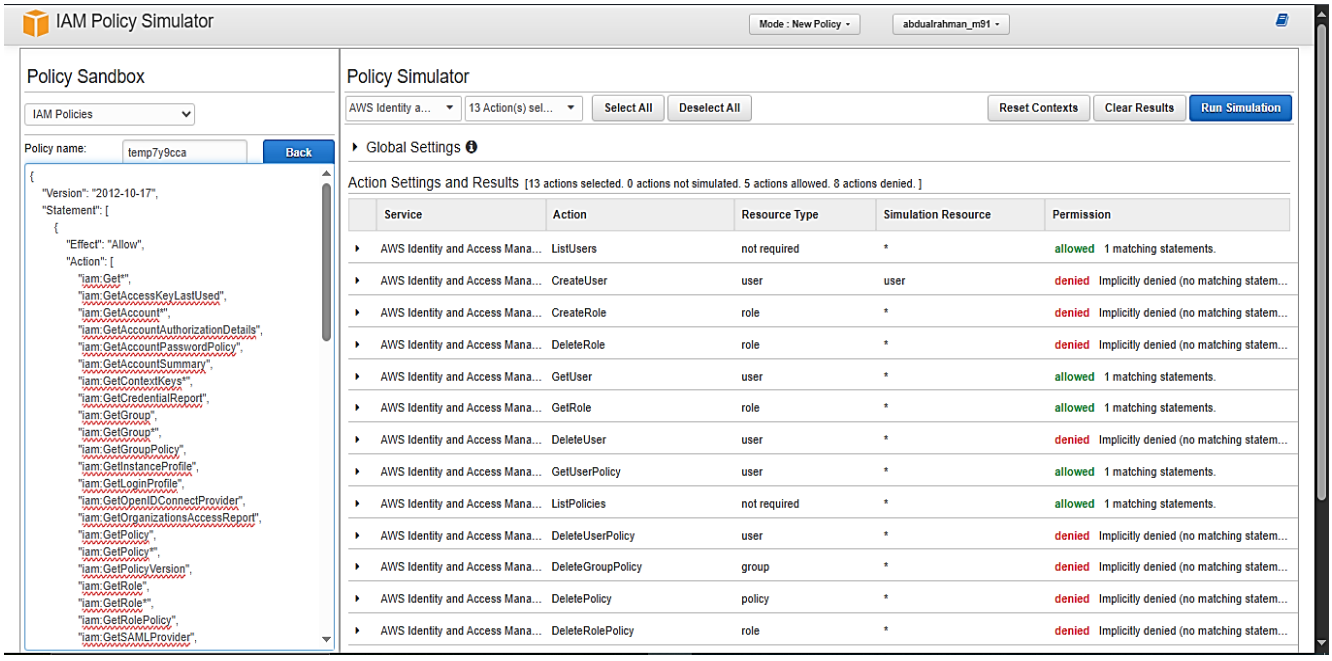

Fig. 7 - IAM Policy Simulator results after PPO-based optimization.

Fig. 6 compared with 7 provides the clue on the practical implications of the proposed framework at the policy implementation level. Before optimization, the IAM policy under consideration allows an excessive scope of operations including authorizations that should not be a part of the regular operations work. The optimized policy, which has gone through optimization, shows that allowed actions are significantly smaller i.e. there is an increased strictness in adherence to the principle of least privilege but the permissions must remain to maintain the continuity

of services. This qualitative test has been conducted using AWS IAM Policy Simulator, where the ability to test the effective permissions as a consequence of deployed policy configurations is possible. The quantitative changes that were seen through the Least Privilege Reduction Score (LPRS) are directly translated into the enforceable IAM policy behavior in the AWS environment. These results assist in highlighting the practical workability and the working relevance of the suggested optimization framework.

Conclusions

The current paper has introduced a new framework of automating the hardening of AWS IAM policy that bridges the gap between adaptive optimization and operational safety. The proposed system, with its combination of the flexibility of the Proximal Policy Optimization (PPO) and deterministic safety constraints, addresses the weaknesses of each of the two aforementioned tools, both of which lack the capability to be adapted through choice and the former, which is not always stable through learning. The serverless, event-driven architecture of the framework guarantees that downsizing of privileges is carried out constantly and effectively in a way that does not interfere with the mission-critical workflows.

Experimental analysis of real world IAM data showed that the framework was successful and that the accuracy of high risk permissions, 96.5% and a reduction of effective privileges, 78.9% were much higher than rule based baselines. Importantly, the inclusion of deterministic guardrails combined with the fact of a 100 per cent. recall of the necessary services effectively nullified the possibility of accidental disruption of service, a significant aspect of this strategy as compared to the other RL-based strategies like those introduced by Saqib et al. [12].

Additionally, our solution is online-executable and scalable as opposed to formal verification methods described in D' Antoni et al. Dantiloni et al., 2024, and can respond to dynamic workloads on the cloud in real-time.

Study Limitations: Although this study has been very promising, it has some limitations. The experimental assessment was made under a controlled AWS setup with a specific account that was set to create the simulation of the realistic organizational workloads. Even though this configuration is safe and reproducible, it might not be sufficient to represent the variability and size of large enterprise cloud infrastructures.

Besides, the analysis was confined to one cloud provider (AWS), which can be a limitation to extending the results to heterogeneous multi-cloud settings.

The future of work is to broaden this architecture to a multi-cloud platform (including Azure and GCP) to provide a coherent cross-platform security architecture. We will also want to explore Federated Learning methods to allow the model training with privacy preservation across various organization, and adversarial training to help the agent to be robust to advanced evasion strategies in the changing threat environment.

References

[1] D'Antoni, L., Ding, S., Goel, A., Ramesh, M., Rungta, N., & Sung, C. (2024). Automatically reducing privilege for access control policies. Proceedings of the ACM on Programming Languages, 8(OOPSLA2).
[2] Amazon Web Services. (2026). IAM Policy Evaluation Logic. Amazon Web Services Documentation.
[3] N2WS. (2025). 49 cloud computing statistics you must know in 2025. N2WS Blog.
[4] Check Point & DuploCloud. (2024). Cloud security report: Misconfigurations and limited visibility plague enterprises. DuploCloud Blog.
[5] Lu, H., Lin, J., Zhou, Y., & Wang, X. (2025). Uncovering cloud access risks under real-world IAM practices. Proceedings on Privacy Enhancing Technologies, 2025(2).
[6] Horizon3.ai. (2023). AWS misconfiguration leads to buckets of data. Horizon3.ai Attack Research.
[7] Eiers, W., Sankaran, G., & Bultan, T. (2023). Quantitative policy repair for access control on the cloud. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA) (pp. 564–575).
[8] National Institute of Standards and Technology (NIST). (2020). Zero Trust Architecture (NIST Special Publication 800-207).
[9] Soveizi, N., & Karastoyanova, D. (2025). Reinforcement learning–driven adaptation chains: A robust framework for multi-cloud workflow security. arXiv preprint arXiv:2501.06305.
[10] Aref, Z., Wei, S., & Mandayam, N. B. (2025). Human–AI collaboration in cloud security: Cognitive hierarchy-driven deep reinforcement learning. arXiv preprint arXiv:2502.16054.
[11] Mahmood, G. S., Hasan, N., Abed, H. N., & Jalil, B. A. (2022). An efficient and secure auditing system of cloud storage based on BLS signature. International Journal of Computing and Digital Systems, 12(7), 1491–1501.
[12] Saqib, M., & Mehta, D. (2025). Adaptive security policy management in cloud environments using reinforcement learning. arXiv preprint arXiv:2505.08837.
[13] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
[14] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.
[15] Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. Proceedings of the International Conference on Learning Representations (ICLR).
[16] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (ICLR).
[17] Amazon Web Services. (2026). IAM Policy Simulator. Amazon Web Services Documentation.