Au-delà des garde-corps : Pourquoi la sécurité de l'infrastructure AI est difficile

C'est le dernier article de notre série en trois parties sur la sécurité de AI.

- **Partie 1 : «** **Ce que personne ne vous dit sur la sécurisation des applications d'AI : Démystifier les barrières de l'AI** **»** a montré comment les pratiques DevSecOps et les modèles de sécurité spécialisés contribuent à sécuriser et aligner les applications LLM.
- **Partie 2 : «** **Automatisation de la politique de sécurité des barrières de sécurité pour une surveillance et un contrôle d'accès** améliorés **»** a examiné les politiques qui alimentent les barrières de sécurité et la manière dont l'analyse assistée par LLM peut aider à les créer et à les gérer.
- **Partie 3 : « Au-delà des garde-corps : Pourquoi la sécurité de l'infrastructure AI est difficile » (*vous êtes là !*) :** Cet article se concentre sur les systèmes qui exécutent des charges de travail d'AI et les contrôles qui assurent la résilience de ces systèmes.

Les contrôles orientés vers les modèles ne répondent qu'à une partie de la question de sécurité : *Comment les politiques peuvent-elles suivre le rythme dans l'ensemble de l'écosystème d'AI ?*

Cet article examine l'infrastructure sous le modèle. Une application d'AI s'exécute sur des réseaux, assume des rôles, lit des secrets, écrit des journaux et se connecte à des bases de données, des magasins d'objets, des files d'attente et des points de terminaison de modèle. Un filtre d'invite ne peut pas limiter un rôle de tâche ECS avec un accès S3 générique ou un groupe de sécurité qui autorise un trafic étendu de charge de travail à charge de travail. Ces risques nécessitent une politique d'infrastructure.

Au-delà de l'invite : Sécuriser l'infrastructure derrière lAI

Les applications d'AI sont des systèmes distribués. Une seule application peut inclure des conteneurs, des fonctions, des files d'attente, des bases de données, des magasins d'objets, des terminaux de modèle, des secrets, des services de

journalisation et plusieurs identités de charge de travail. Chaque composant ajoute un chemin réseau, une identité, une autorisation ou une configuration qui peut étendre le rayon d'explosion de l'application.

Ici, le moindre privilège couvre quatre zones connectées :

- **Réseau :** Autoriser uniquement les sources, destinations, protocoles et ports requis pour le trafic nord-sud entrant ou sortant de l'application et le trafic est-ouest entre les charges de travail
- **Authentification :** Donnez à chaque charge de travail une identité spécifique et vérifiable et une relation de confiance étroite, en utilisant des identifiants gérés plutôt que des secrets partagés ou statiques
- **Autorisation :** Permettre à cette identité authentifiée d'effectuer uniquement les actions requises sur des ressources, des API et des données cloud spécifiques plutôt que d'accorder un accès générique
- **Renforcement de la configuration :** Appliquer des paramètres par défaut sécurisés, tels que le protocole TLS, le chiffrement, la journalisation des audits et la segmentation, qui réduisent l'exposition et rendent les autres contrôles plus difficiles à contourner

Ces contrôles dépendent les uns des autres. Une politique IAM étroite laisse encore un risque si la charge de travail a une large accessibilité réseau. Une règle réseau précise est incomplète si la charge de travail partage des informations d'identification ou s'exécute avec des paramètres par défaut non sécurisés.

Pour réduire l'accès en toute sécurité, les équipes de sécurité ont besoin d'une limite d'application actuelle, de preuves de ses dépendances réelles et de contrôles minimums dans les quatre domaines. Ils doivent également effectuer ces changements sans interrompre la production.

Pourquoi le moindre privilège est difficile

« Ne donnez à chaque charge de travail que la connectivité et les autorisations dont elle a besoin » est facile à déclarer. Cependant, la limite de l'application et son accès requis ne sont pas stockés dans un seul endroit faisant autorité.

Le rapport Microsoft sur l'état des autorisations cloud a révélé que les identités des charges de travail étaient plus nombreuses que les identités humaines 10 pour 1. En moyenne, 80 % des identités de charges de travail sont inactives et utilisent moins de 5 % des autorisations qui leur sont accordées. Cela laisse un vaste vivier d'accès aux machines qui n'est peut-être plus nécessaire, mais qui peut encore

étendre l’impact d’une compromission.

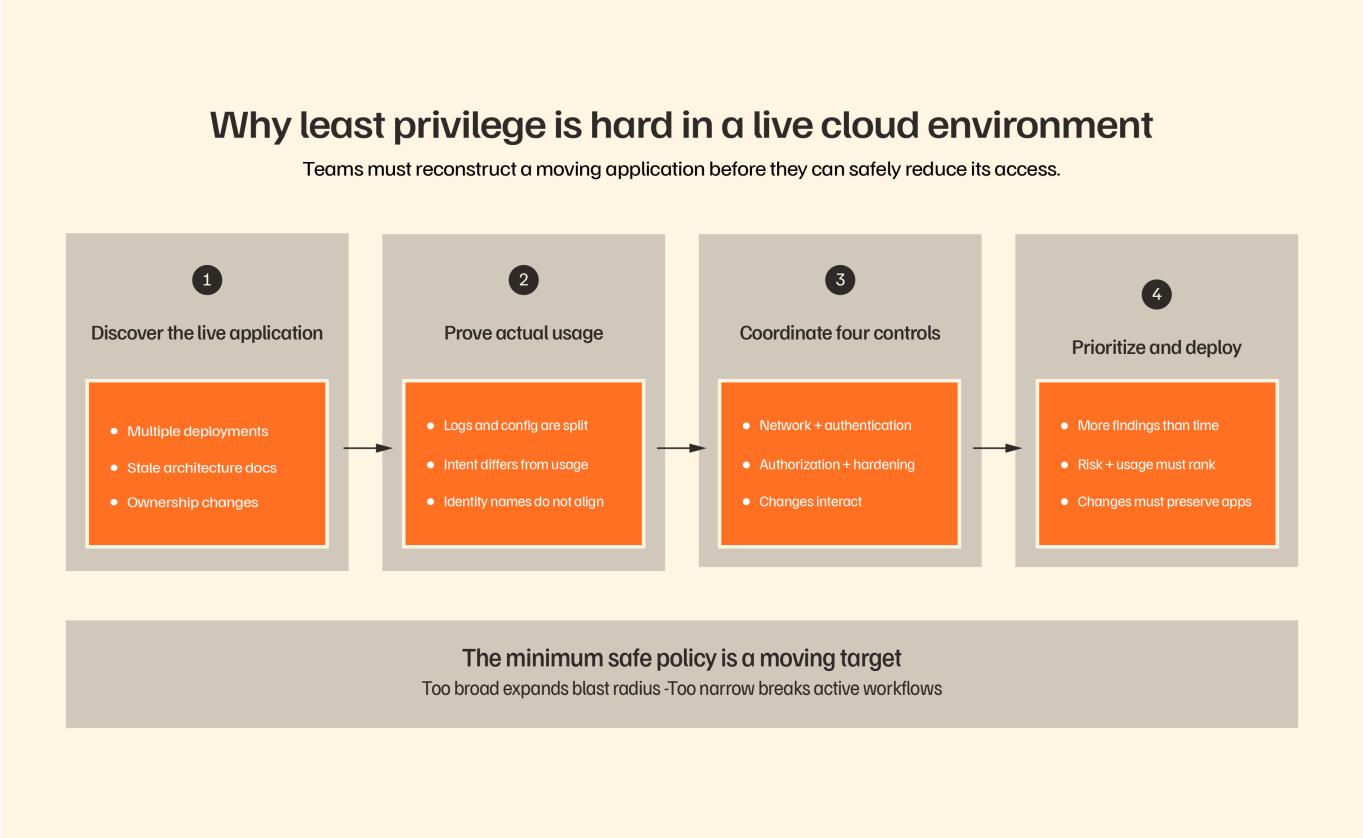


Figure 1 : Le moindre privilège est difficile, car les équipes doivent reconstruire une application changeante, prouver quel accès est actif, coordonner quatre zones de contrôle et prioriser les changements sans rompre les flux de travail valides.

Classement	Problématique	Importance
1	L’accès requis diffère de l’accès provisionné	Terraform et la configuration indiquent ce qui a été provisionné ; CloudTrail, les journaux de flux VPC et les journaux de charges de travail indiquent ce qui a été réellement utilisé. Aucune vue n’est complète seule

Classement	Problématique	Importance
2	L’application en direct est difficile à reconstruire	Les ressources peuvent partager un compte ou un VPC sans appartenir à la même application, tandis qu’une seule application peut couvrir les services et les déploiements. L’architecture, les flux métier et la propriété changent également, de sorte que la documentation seule ne peut pas montrer si une ressource ou un chemin d’accès est toujours actif
3	L’accès effectif à l’identité est fragmenté	Les rôles apparaissent différemment sur Terraform, ECS, les politiques de ressources et CloudTrail, tandis que la fédération et la chaîne de rôles peuvent masquer l’appelant d’origine. Les autorisations effectives découlent également des politiques d’identité et de ressources, des limites, des politiques de session, des politiques de contrôle des services et des refus explicites ; aucune politique ne révèle l’accès réel
4	Le moindre privilège couvre plusieurs couches de contrôle	La restriction de l’IAM ne supprime pas l’accessibilité du réseau et une règle réseau précise ne corrige pas les informations d’identification partagées ou la configuration non sécurisée
5	Les recommandations doivent être priorisées et déployables	Les équipes de sécurité reçoivent plus de résultats qu’elles ne peuvent en traiter en même temps. Ils ont besoin d’un contexte d’utilisation et de risque pour distinguer les expositions à fort impact du nettoyage de priorité inférieure, puis d’une politique sous forme de code qu’ils peuvent valider dans les flux de travail d’ingénierie existants

Les autorisations excessives deviennent plus dangereuses lorsqu’elles se chevauchent avec d’autres faiblesses. Un rapport 2024 de Tenable sur les risques

liés au cloud a révélé que 38 % des organisations disposaient de charges de travail cloud combinant exposition publique, vulnérabilité critique et privilèges excessifs, une combinaison appelée « trilogie du cloud toxique ». Le moindre privilège ne supprime pas la vulnérabilité, mais peut réduire ce qu'un attaquant peut atteindre ou faire après l'avoir exploitée.

Les documents d'architecture sont des snapshots utiles, et non des preuves du comportement actuel. Les déploiements se multiplient, les flux de travail changent et la propriété passe d'une équipe à l'autre. Les personnes qui sécurisent un environnement peuvent ne pas savoir si une ressource, une autorisation ou un chemin réseau prend toujours en charge un flux de travail actif.

Les journaux d'utilisation comblent une partie de cette lacune. CloudTrail indique quel principal a effectué une action API sur une ressource et quand. Les journaux de flux VPC indiquent les points de terminaison communiqués, les ports et la fréquence. Les journaux de charge de travail et de service renforcent le comportement des applications. La corrélation de ces sources permet de séparer l'accès requis des autorisations et des chemins réseau étendus, obsolètes ou inutilisés. Elle fournit également aux équipes de sécurité des preuves pour décider quels changements réduisent le plus les risques.

La portée de l'identité contrôle également le rayon de projection. Dans le cadre d'engagements Mandiant impliquant des environnements cloud et SaaS majeurs au cours du deuxième semestre 2025, Google Cloud a signalé que les acteurs malveillants exploitaient les problèmes d'identité pour un accès initial dans 83 % des incidents, tandis que les données étaient ciblées dans 73 % des incidents. Lorsqu'une identité est compromise, le moindre privilège détermine la portée de cet accès.

Ensuite : Comment ASTRA transforme la télémétrie en un privilège minimal

Chez Everpure, les équipes de sécurité et d'ingénierie de fiabilité du site (SRE) utilisent ASTRA (Automated Service Topology and Risk Analysis) pour sécuriser notre infrastructure. ASTRA associe la configuration de l'infrastructure à l'utilisation observée pour reconstruire les topologies d'applications, identifier les accès étendus, prioriser les risques les moins privilégiés et générer des recommandations de stratégie en tant que code.

Notre prochaine série de blogs sur ASTRA répondra à trois questions :

- Comment ASTRA déduit-elle la topologie de votre application ?
- Comment ASTRA détecte-t-elle les lacunes liées au renforcement de la configuration ?
- Comment ASTRA priorise-t-elle les recommandations de moindre privilège ?

ARGOS, notre moteur de règles d'AI, en sera l'exemple. La série suivra l'ASTR de la télémétrie à la topologie des applications et aux flux observés, puis au risque priorisé et au code Terraform le moins privilégié.