TRUST

De la confiance déclarée à l'autorité vérifiable

Identité, délégation et pouvoir de décision dans un système agentique

Comment Sawbill transforme une identité vérifiée en autorité précise, limitée, révocable et liée à une action.

LECTEURS

Architectes / Responsables sécurité / Responsables de plateforme IA / Responsables risque et gouvernance

Sommaire

De la confiance déclarée à l'autorité vérifiable

L'enjeu en une minute

Lorsqu'un agent se présente comme « agent de développement », « reviewer » ou « administrateur », ce label ne prouve rien. Même une identité correctement authentifiée ne répond pas à la question décisive : cette identité a-t-elle le droit d'accomplir **cet acte**, sur **cet objet**, dans **ce contexte**, à **cet instant** ?

Sawbill remplace la confiance implicite par une chaîne vérifiable :

> **un credential est vérifié, une identité est résolue, une autorité est évaluée, puis son usage est lié à une demande exacte.**

Le résultat n'est pas un rôle global ni un jeton réutilisable. C'est un contexte d'autorisation borné, traçable et invalidable.

Pourquoi les contrôles habituels ne suffisent plus

Les systèmes traditionnels s'appuient souvent sur quatre raccourcis :

- le compte connecté est supposé représenter la bonne personne ;
- le rôle `admin` est supposé couvrir toutes les décisions nécessaires ;
- le worker qui exécute une tâche est supposé agir au nom du bon demandeur ;
- une signature valide est supposée valoir approbation.

Ces raccourcis deviennent dangereux avec des agents capables de déléguer, de changer de modèle ou de runtime, de travailler sur plusieurs dépôts et de solliciter des effets externes. Un compte peut être compromis. Un agent peut être relancé sous une autre identité de workload. Une autorisation valide pour un dépôt ne l'est pas nécessairement pour une infrastructure de production.

Sawbill traite donc séparément six objets que les applications confondent souvent :

Objet	Question à laquelle il répond	Ce qu'il ne prouve pas
Credential	Quel matériel d'authentification est présenté ?	L'identité ou le droit d'agir
Identité vérifiée	À quel principal ce matériel est-il lié ?	Un rôle ou une permission
Contexte de confiance	Quelles ancres, politiques et données de révocation ont été utilisées ?	Une décision métier
Autorité accordée	Quels actes sont permis sur quels sujets ?	Que l'acte est admissible maintenant
Délégation	Quelle part bornée d'une autorité parente est transmise ?	Un élargissement de pouvoir
Contexte d'autorisation	Cet usage exact est-il permis ?	Le démarrage ou le succès de l'action

Cette séparation est le fondement du modèle de confiance Sawbill.

La chaîne d'autorité Sawbill

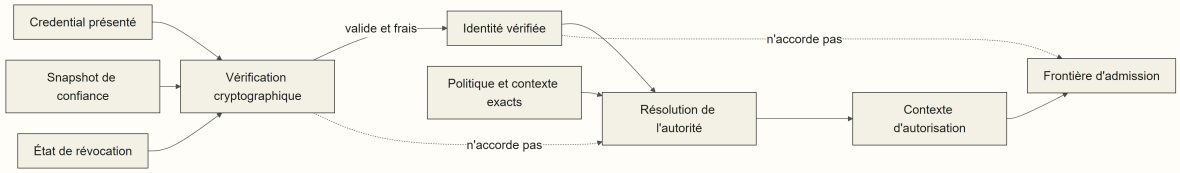

Figure 1 - Sawbill

1. Vérifier le credential

Sawbill commence par vérifier le matériel effectivement présenté : certificat, assertion d'identité, credential de workload ou autre mécanisme admis par le profil. La vérification lie notamment :

- le format et la version attendus ;
- la chaîne et l'ancre de confiance ;
- l'usage prévu du credential ;
- la période de validité ;
- les informations disponibles de révocation ;
- l'instant et le contexte de vérification.

Une chaîne inconnue, un contrôle de révocation indisponible ou un profil ambigu ne devient pas « probablement valide ». Le résultat reste bloqué ou inconnu.

2. Résoudre une identité stable

Le credential est ensuite relié à une identité Sawbill stable. Cette identité n'est ni une adresse courriel, ni un nom affiché, ni un ID de session. Les credentials peuvent tourner sans changer le principal ; inversement, deux credentials portant un même nom ne sont pas automatiquement équivalents.

Pour les workloads, Sawbill peut s'appuyer sur des identités courtes et rotatives, comme les SVID définis par SPIFFE. SPIFFE fournit une identité vérifiable de workload et des domaines de confiance. Sawbill conserve la décision métier qui détermine ce que ce workload est autorisé à faire.

3. Évaluer l'autorité

Une autorité Sawbill décrit une relation explicite :

- le principal autorisé ;
- les actions permises ;
- les sujets ou ressources concernés ;
- les contraintes de portée, de temps ou de projet ;
- les limites de délégation ;
- la provenance de l'autorité ;
- la politique qui gouverne son interprétation.

La résolution confronte ces données à la demande réelle. Elle ne demande pas simplement si le principal « possède le rôle développeur », mais s'il peut, par exemple, proposer une modification sur le service d'export du projet A pendant une tentative donnée.

4. Lier l'autorisation à un usage exact

Le résultat est un contexte d'autorisation lié à l'action, au sujet, au contrat, à la tentative, à l'audience et au moment de la décision. Son usage reste borné à la finalité pour laquelle il a été produit.

Ainsi, copier le contexte vers une autre tentative ou une autre ressource ne fonctionne pas, même si le même agent et le même credential sont présents.

Où Sawbill s'insère dans l'identité existante

Sawbill ne remplace ni l'annuaire, ni l'autorité de certification, ni le service d'identité de workload. Ces systèmes continuent d'authentifier leurs identités dans leur périmètre. Sawbill ajoute la décision qui manque à une chaîne agentique : quel acte exact cette identité peut-elle demander maintenant ?

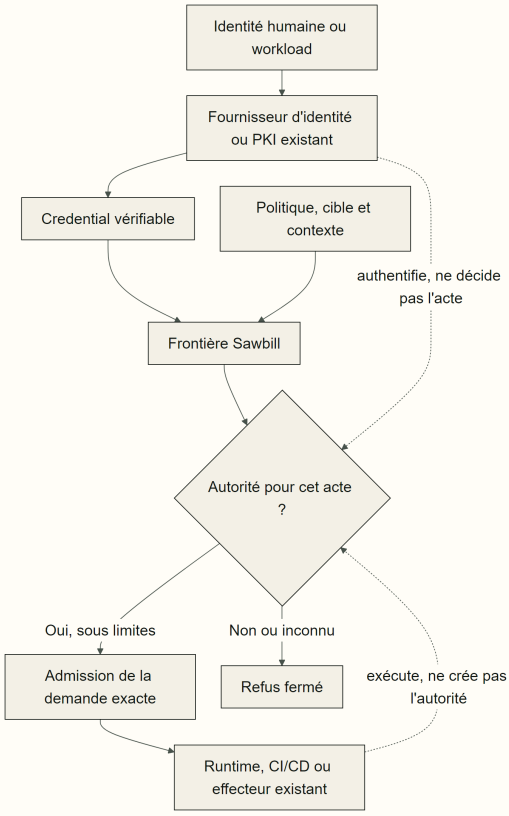

Figure 2 - Sawbill

Les standards d'identité de workload, les certificats X.509 et les mécanismes d'authentification restent donc des points d'intégration. Ils apportent une identité vérifiable ; Sawbill conserve l'autorité métier, la délégation, la fraîcheur et l'admission. La frontière de confiance se déplace du compte ou du runtime vers une décision explicite, liée à la demande.

Un certificat identifie ; Sawbill autorise

Sawbill s'appuie sur la PKI sans lui déléguer toute sa gouvernance. Le profil X.509 suit les fondations de validation de chemin et de certificats de la RFC 5280. Un certificat propriétaire peut lier une clé publique à une identité, une installation et une portée. Il ne devient jamais, à lui seul, un droit de modifier une baseline, d'accepter un livrable ou d'effectuer un déploiement.

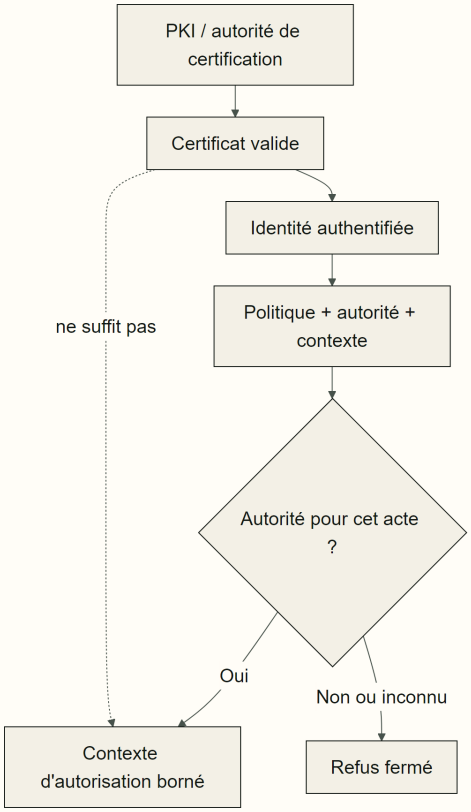

Figure 3 - Sawbill

Cette architecture accepte une chaîne de confiance choisie et administrée par l'organisation. La clé privée du propriétaire reste sous son contrôle et n'est pas remise aux agents. L'émission et le renouvellement peuvent rester assurés par l'infrastructure existante ; le service de certificats ne décide jamais de l'autorité Sawbill.

Déléguer sans élargir

La délégation est indispensable aux systèmes multi-agents : un orchestrateur confie une analyse à un spécialiste, qui peut lui-même demander une vérification. Mais une délégation naïve transforme rapidement une permission limitée en pouvoir transitif.

Sawbill impose une règle simple : une délégation enfant doit être une atténuation mécaniquement vérifiable de son parent.

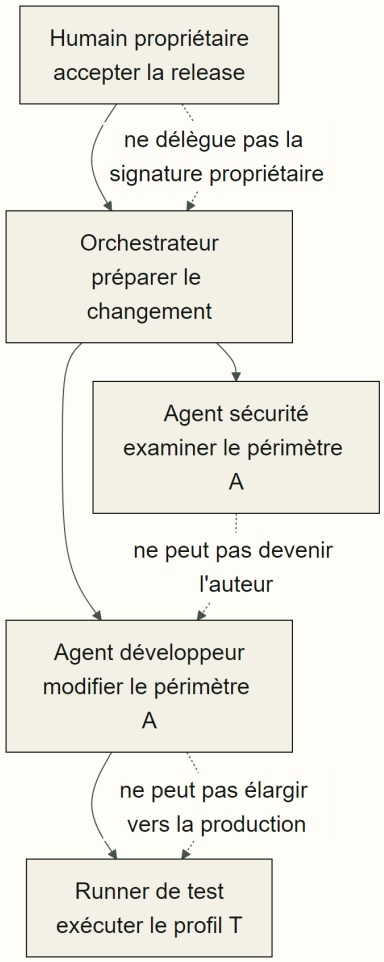

Figure 4 - Sawbill

Chaque enfant conserve une relation explicite avec son parent, une profondeur, une portée et des contraintes au moins aussi strictes. Une valeur incomparable ou ambiguë bloque la délégation. La révocation ou l'expiration d'un ancêtre rend la descendance inutilisable selon la politique applicable.

La séparation des devoirs porte sur le contrôle réel

Deux noms d'agents différents ne prouvent pas l'indépendance. Ils peuvent partager le même modèle, le même compte fournisseur, la même clé, le même runtime ou le même contrôleur humain.

Sawbill peut évaluer la séparation des devoirs selon plusieurs axes :

- principal et principal de contrôle ;
- organisation ou équipe ;
- credential et domaine de clés ;
- workload et environnement d'exécution ;
- modèle et profil de modèle ;
- relation entre auteur, validateur et approbateur.

L'indépendance reste proportionnée au risque. Un contrôle mécanique peut vérifier que deux credentials et deux runtimes diffèrent ; il ne prétend pas démontrer l'absence de collusion organisationnelle sans information appropriée.

Scénario : autoriser une fonction d'export de données

Une responsable produit demande l'ajout d'un export de données clients. Le travail touche une API, un dépôt d'infrastructure et une politique de résidence.

1. Son credential est vérifié sous le profil de l'organisation.
2. Son identité Sawbill est résolue indépendamment de son compte applicatif.
3. Son autorité permet la création de l'intention, mais pas son approbation sécurité ni son déploiement en production.
4. L'orchestrateur reçoit une délégation limitée à la préparation du changement.
5. L'agent développeur peut modifier le périmètre défini ; il ne peut pas élargir sa mission au système de facturation voisin.
6. Le reviewer sécurité est sélectionné sous les critères d'indépendance requis.
7. L'approbation finale exige l'autorité et, selon le profil, le quorum prévus pour la version exacte du livrable.

Le système ne repose jamais sur l'idée que « l'agent travaille pour la bonne équipe ». Chaque transition importante possède son autorité propre.

Ce que cette approche change pour l'entreprise

Pour la sécurité

- réduction des permissions permanentes confiées aux agents et workers ;
- révocation et rotation visibles dans les décisions courantes ;
- refus des escalades fondées sur des comptes, rôles ou labels ambigus ;
- séparation vérifiable entre identité, autorité et exécution.

Pour l'ingénierie

- délégation multi-agent sans perte de traçabilité ;
- politiques réutilisables entre interfaces et environnements ;
- décisions reproductibles à partir des mêmes entrées ;
- intégration de plusieurs fournisseurs d'identité sans fusion des sémantiques.

Pour le risque et l'audit

- attribution de chaque décision à un principal vérifié ;
- portée et limites d'une autorité explicites ;
- historique des rotations, révocations et délégations ;
- distinction claire entre authentification et approbation.

Ce que Sawbill affirme - et ce qu'il n'affirme pas

Sawbill peut établir	Sawbill ne déduit pas automatiquement
Un credential a été vérifié sous un profil et un contexte précis	Que toute affirmation portée par ce credential est vraie
Une identité a été résolue à un cutoff donné	Que cette identité possède un droit d'agir
Une autorité permet une action bornée	Que l'action doit être exécutée ou acceptée

Sawbill peut établir	Sawbill ne déduit pas automatiquement
Une délégation atténue le pouvoir de son parent	Que les acteurs sont organisationnellement indépendants en toute circonstance
Une décision d'autorité a été prise sur des entrées exactes	Que cette décision restera utilisable après révocation ou changement de politique

Cette précision n'affaiblit pas la confiance : elle la rend vérifiable.

Références officielles

Source	Statut externe	Apport utilisé	Limite conservée par Sawbill
RFC 5280 - PKIX Certificate and CRL Profile	IETF Proposed Standard, mise à jour par plusieurs RFC	Certificats X.509, extensions, validation de chemin et CRL	Un certificat n'est pas une autorisation Sawbill
NIST SP 800-63-4 - Digital Identity Guidelines	Version finale publiée en juillet 2025	Vocabulaire d'assurance pour preuve d'identité, authentification et fédération	Les niveaux d'assurance ne deviennent pas des permissions métier
SPIFFE Standard	Spécification ouverte maintenue ; Workload API déclarée stable	Identité portable de workloads, SVID et domaines de confiance	Une identité SPIFFE authentifie un workload ; Sawbill décide de son autorité
NIST SP 800-207 - Zero Trust Architecture	Publication NIST finale	Absence de confiance implicite fondée sur le réseau ou la propriété	Sawbill applique ce principe aux actes agentiques et aux objets gouvernés

Les références externes restent normatives dans leur propre périmètre. Leur présence ne constitue ni une certification de Sawbill, ni une garantie juridique, ni une équivalence avec un niveau d'assurance non évalué.