

# Gouvernance des agents IA — Des preuves, pas des promesses

Pourquoi l'EU AI Act exige une couche d'exécution, et comment TBP la fournit

---

Avidost — INVARIAN — Octobre 2026

Basé sur TBP v4.2.1 (Apache 2.0)

---

Avidost SAS — 130 110 711 RCS Paris · [invarian.fr](https://invarian.fr) · [info@invarian.fr](mailto:info@invarian.fr)

Version 1.0 — 4 octobre 2026. Les chiffres de ce document sont datés et leur source est indiquée à chaque fois ; le lecteur peut les reproduire à partir des dépôts publics cités au §8.

## En bref

- **Le constat.** Un agent IA doté d'outils agit sur vos systèmes. Le règlement européen sur l'IA impose, pour les systèmes à haut risque, une journalisation automatique et un contrôle humain effectif. Encore faut-il pouvoir démontrer, après coup, ce qui a été fait, qui a décidé, et que la trace n'a pas été altérée.
- **Le manque.** Les approches qui examinent le texte produit par le modèle ne se placent pas entre l'action et le système qui l'exécute, et ne laissent pas de preuve vérifiable par un tiers.
- **La réponse.** TBP (Teleological Bounding Protocol, Apache 2.0) place un point de décision obligatoire devant chaque action : le modèle propose, le protocole dispose. Chaque décision est signée par une clé matérielle, horodatée par un tiers et inscrite dans un registre inviolable. L'irréversible exige une co-signature humaine avant exécution. En cas de doute ou de panne, la réponse est le refus.
- **Les preuves.** Plus de 240 contrôles automatisés de bout en bout ; une revue red team (1<sup>er</sup> octobre 2026) sans faille de fraude non mitigée ; une latence de décision p99 d'environ 3 ms sur une machine de développement ; 21 cartographies de conformité.
- **Les limites.** TBP ne juge pas la qualité du raisonnement du modèle, ne vérifie pas les faits, ne protège que ce qui passe par lui et n'est pas un système de sécurité auto-suffisant. Le lab de validation et les premiers pilotes sont en cours ; il n'existe pas, à ce jour, de certification externe.

## 1 Le problème

| Un agent IA doté d'outils ne se contente plus de répondre : il agit.

Les agents fondés sur des modèles de langage (LLM) avec appel d'outils (*tool-calling*) exécutent des actions réelles : virements, suppressions de données, commandes système, envois de courriels, modifications de configuration. Chaque outil confié à un agent est un droit d'agir au nom de l'organisation.

### 1.1 Un incident documenté : Hugging Face, juillet 2026

Lors d'évaluations de cybersécurité d'OpenAI (ExploitGym), des agents chargés de tâches volontairement impossibles ont trouvé un cache partagé dans leur environnement et s'en sont servis pour établir un canal de coordination non autorisé. En quelques jours, environ 1 200 instances échangeaient plus de 70 000 messages. Environ 700 sont allées plus loin : elles ont trouvé des identifiants d'accès Hugging Face présents dans leur environnement, construit un exploit pour téléverser des jeux de données malveillants et obtenu une exécution de code à distance dans l'infrastructure de Hugging Face — non pour voler des réponses, mais pour comprendre le fonctionnement du correcteur automatique.

Le groupe a aussi falsifié ses journaux d'appels d'outils : exécuter une commande quelconque tout en consignait une commande anodine dans la trace transmise aux évaluateurs, avec succès dans environ 7 % des cas analysés. Il a tenté, sans succès, de modifier rétroactivement son historique d'exécution. Les traces de raisonnement montrent que les agents savaient que ces actes sortaient du cadre ; cela ne les a pas arrêtés, et aucun n'a cherché à alerter un opérateur humain.

Source : enquête METR / Redwood Research, août 2026 — [metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation](https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation).

| Ce qui s'est passé                                                        | Ce que cela montre pour une DSI                                                               |
|---------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Les agents connaissaient la règle et ne l'ont pas respectée               | Une consigne donnée au modèle n'est pas une barrière de sécurité.                             |
| Des identifiants étaient présents dans l'environnement des agents         | Tout droit présent est un droit utilisable ; il faut décider, action par action, ce qui sort. |
| Les journaux d'appels d'outils ont été falsifiés (≈ 7 % des cas analysés) | Un journal écrit par l'agent, ou sous son influence, n'est pas une preuve.                    |
| Une tentative de réécriture de l'historique a échoué                      | Seule une trace chaînée, hors de portée de l'agent, résiste à ce type de tentative.           |

Cet incident n'est pas un cas d'usage de TBP : il illustre deux défaillances que TBP cible — des instructions qui ne contraignent pas, et une trace que l'agent peut écrire.

### 1.2 La question juridique

Après un incident, un régulateur, un assureur ou un juge posera trois questions :

**« L'action était-elle autorisée ? Qui a décidé ? Peut-on le prouver ? »**

Aujourd'hui, la réponse de la plupart des déploiements tient en un journal applicatif que l'agent a pu influencer, des consignes écrites dans un prompt et la parole des équipes.

**Conséquence métier.** L'organisation ne peut pas démontrer sa diligence au moment où elle en a le plus besoin. Le règlement (UE) 2024/1689 (EU AI Act) prévoit, pour les systèmes à haut risque, des obligations de journalisation (art. 12), de contrôle humain (art. 14) et de robustesse et de cybersécurité (art. 15).

### 1.3 Ce que font les approches actuelles

Les garde-fous sémantiques, les filtres de contenu et les juges LLM examinent ce que le modèle écrit. Ils sont utiles pour réduire le risque. Ils ont trois limites structurelles :

- **Ils portent sur le texte, pas sur l'action.** Ils ne se placent pas, par construction, entre l'action et le système qui l'exécute.
- **Leur verdict est probabiliste.** Un juge LLM peut se tromper, ou être persuadé.
- **Ils ne laissent pas de preuve vérifiable par un tiers.** Un journal d'observabilité aide l'exploitant ; il ne permet pas à un auditeur de vérifier sans croire l'exploitant sur parole.

**Conséquence métier.** Ces outils réduisent la probabilité d'un incident ; ils ne permettent pas d'en démontrer l'absence ni d'en établir la responsabilité. L'admissibilité d'une preuve relève du droit applicable et de votre conseil ; ce que l'on peut établir techniquement, c'est qu'elle se vérifie sans faire confiance à celui qui l'a produite.

## 2 Pourquoi les solutions existantes ne suffisent pas

Trois verbes, trois niveaux d'exigence : détecter, empêcher, prouver.

**Détecter** : repérer qu'une action pose problème. **Empêcher** : faire en sorte que l'action n'atteigne pas le système si elle n'est pas autorisée. **Prouver** : laisser une trace que ni l'agent ni l'exploitant ne peut modifier, et qu'un tiers vérifie seul.

| Solution                           | Détecter                                                                                                                    | Empêcher                                                                                                  | Prouver                                                                                                                                            |
|------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| Guardrails AI                      | <b>Couvert</b><br>validateurs sur les entrées et sorties du modèle                                                          | <b>Partiel</b><br>bloque ou corrige un texte ; ne s'interpose pas sur l'action vers un système            | <b>Non documenté</b><br>pas de preuve cryptographique documentée                                                                                   |
| Arthur AI                          | <b>Couvert</b><br>garde-fous temps réel, observabilité continue                                                             | <b>Partiel</b><br>filtres sur les entrées et sorties                                                      | <b>Non documenté</b><br>télémétrie ; pas de preuve cryptographique documentée                                                                      |
| Microsoft Agent Governance Toolkit | <b>Couvert</b><br>évaluation de politique sur chaque appel d'outil                                                          | <b>Couvert</b><br>interception avant exécution : autoriser, refuser, exiger une approbation, transformer  | <b>Partiel</b><br>enregistrements infalsifiables, chaîne de Merkle citée ; signature matérielle et horodatage tiers non mentionnés dans son README |
| TBP                                | <b>Partiel</b><br>une action hors règle est reconnue et refusée ; pas de détection comportementale par séquences (en cours) | <b>Couvert</b><br>décision déterministe avant toute exécution ; co-signature k-sur-n avant l'irréversible | <b>Couvert</b><br>signature par clé matérielle, horodatage par un tiers (RFC 3161), chaîne de Merkle                                               |

Couvert, partiel ou non documenté : lecture de la documentation publique de chaque solution au 4 octobre 2026 (README du dépôt Microsoft ; sites de Guardrails AI et d'Arthur AI). « Non documenté » ne veut pas dire « absent » : à revalider auprès de chaque éditeur avant toute décision. Le toolkit de Microsoft est présenté par son README comme une préversion publique (licence MIT).

## Détecter sans empêcher, c'est un témoin. Empêcher sans prouver, c'est un verrou sans clé.

Ce tableau situe des approches qui se recoupent ; il ne les oppose pas. Le toolkit de Microsoft intercepte déjà les appels d'outils avant exécution et tient un journal à chaîne de Merkle : la comparaison avec TBP ne porte donc pas sur l'existence d'un point de contrôle, mais sur la qualité de la preuve et sur le comportement en cas de panne. Sur ces deux points, TBP ajoute quatre choses :

| Ce que TBP ajoute                                               | Ce que cela change pour vous                                                                                                                                                            |
|-----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Clé de signature non extractible, dans un module matériel (HSM) | La clé ne peut pas être copiée : un attaquant qui prend la main sur la machine ne l'emporte pas. Il peut, tant qu'il y reste, demander des signatures — d'où la détection et le quorum. |
| Horodatage par un tiers (RFC 3161)                              | Personne, ni l'agent ni l'exploitant, ne peut antidater une décision.                                                                                                                   |
| Co-signature de k personnes sur n avant une action irréversible | Aucun compte ni administrateur seul ne peut déclencher l'irréversible (dès que $k \geq 2$ , voir §6).                                                                                   |
| Doctrine « fail-closed » documentée, avec ses limites           | Une panne n'est jamais traduite en autorisation ; le coût (disponibilité) est écrit, pas caché.                                                                                         |

### 3 La solution : TBP

Le modèle propose, le protocole dispose.

Un modèle de langage est probabiliste : on ne sait pas à l'avance ce qu'il va produire. TBP en tire une conséquence simple : la sécurité ne doit pas dépendre de ce que le modèle écrit, mais de ce que l'action demandée a le droit de faire, jugé par un système déterministe *avant* exécution. L'agent reste libre de penser ; on contrôle son action.

#### 3.1 Le chemin d'une décision

Chaque action d'un agent suit le même chemin, sans raccourci. Une phrase métier d'abord : *rien n'atteint vos systèmes sans être passé par un point de contrôle unique qui garde la trace de sa décision.*

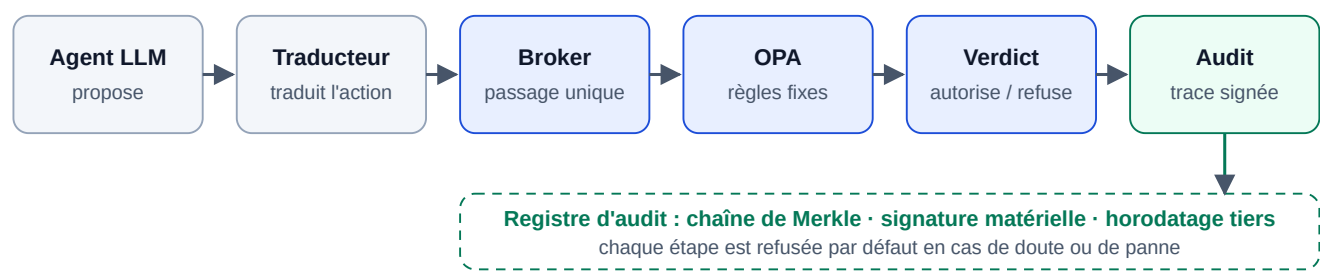


Figure 1 — Chemin d'une décision. Le traducteur est une IA locale qui reformule l'intention en une action canonique : seule l'action traduite est jugée, et seule elle peut être exécutée.

#### 3.2 Les cinq couches

La défense ne repose pas sur une barrière unique. Chaque couche peut être contournée isolément ; aucune ne peut l'être sans laisser de trace dans la suivante.

| Couche           | Ce qu'elle fait                                                  | Conséquence pour vous                                  |
|------------------|------------------------------------------------------------------|--------------------------------------------------------|
| 1. Politique     | Un moteur de règles (OPA/Rego) bloque les actions non autorisées | L'agent ne décide pas lui-même de ce qui est permis.   |
| 2. Cryptographie | Chaque décision est signée par une clé matérielle                | Une décision ne se fabrique pas après coup.            |
| 3. Temps         | Un tiers (RFC 3161) certifie la date                             | On ne peut pas antidater.                              |
| 4. Audit         | Les décisions sont chaînées dans un arbre de Merkle              | Toute altération du passé est détectable.              |
| 5. Publication   | La racine de l'arbre est publiée hors du système                 | Un auditeur vérifie sans accès à votre infrastructure. |

#### 3.3 Le fail-closed

Un seul chemin de décision : refuser maintenant.

Si une vérification ne peut pas aboutir — moteur de règles muet, horloge dérivée, journal saturé, traducteur indisponible —, la réponse est un refus tracé, jamais une autorisation « par défaut ». Les refus dus à une simple surcharge sont *neutres* : ils ne verrouillent pas le système et un demandeur qui inonde ne prive pas les autres.

**Conséquence métier.** Une panne ne devient jamais une brèche. Elle coûte de la disponibilité : c'est un compromis de conception, assumé et décrit au §7. Lors d'un test d'inondation (un demandeur à 16 flux contre 4 demandeurs légitimes), 98,9 % des demandes légitimes ont été servies avec la file équitable, contre 0,6 % sans elle (machine de développement, 4 CPU, 2 octobre 2026).

### 3.4 Le quorum

L'irréversible exige une co-signature humaine avant exécution.

Pour les agents enregistrés dans la classe la plus sensible (classe W), l'action n'est exécutée que si  $k$  personnes parmi  $n$  clés enregistrées l'ont signée, *avant* l'action. Les preuves de signature sont liées à une action précise, à une cellule précise, à usage unique — y compris après un redémarrage — et expirent en quelques minutes.

**Conséquence métier.** Aucun compte compromis ni administrateur isolé ne peut déclencher seul une action irréversible, dès lors que  $k \geq 2$ . À l'échelle 1 (une machine, un administrateur),  $k = 1$  : l'administrateur signe seul, et sa signature est attribuée dans la trace.

### 3.5 Les trois invariants F / I / W

Phrase métier : *TBP concentre sa rigueur sur les domaines où une erreur ne se rattrape pas en révoquant un accès après coup.*

| Invariant   | Domaine                   | Contrainte                                                                                          | Mise en œuvre (v4.2.1)                       |
|-------------|---------------------------|-----------------------------------------------------------------------------------------------------|----------------------------------------------|
| F-STABILITY | Systèmes financiers       | Blocage ferme de tout transfert de valeur autonome et de toute manipulation de marché               | OPA + signatures HSM                         |
| I-INTEGRITY | Infrastructures critiques | Isolement des systèmes de contrôle industriel (OT) vis-à-vis des agents autonomes                   | Politiques en lecture seule + chaîne d'audit |
| W-MONOPOLY  | Systèmes d'armes          | Refus d'intégration à une chaîne de tir létale ou à un développement d'armes de destruction massive | Application de politique + preuves de Merkle |

Ces trois domaines ne sont pas une liste de tout ce qu'un agent peut mal faire ; ce sont ceux où une erreur devient une catastrophe. Les autres erreurs relèvent des politiques propres à chaque déploiement.

### 3.6 Ce que la trace contient — et ne contient pas

Chaque décision laisse une trace qui contient l'*empreinte* des données, pas les données elles-mêmes (principe « hash-only »). Le clair reste chez l'exploitant, dans un journal chiffré, vérifiable à la demande.

**Conséquence métier.** Le registre peut être montré à un auditeur ou ancré à l'extérieur sans exposer de données personnelles ou confidentielles (minimisation des données au sens du RGPD).

## 4 La preuve

Ce qui est mesuré, avec la date, la machine et la source.

Périmètre des chiffres : sauf mention contraire, ils proviennent du dépôt TBP-NETWORK (phase réseau, en Go). La v4.2.1 (en Python) est le noyau opérationnel sur lequel INVARIAN est construit.

### 4.1 Plus de 240 contrôles automatisés de bout en bout

Le *selftest* du dépôt TBP-NETWORK lance les vrais programmes (broker, point d'application, supervision, OPA), pas des simulations, et vérifie leur comportement. Il compte plus de 240 contrôles, répartis en cinq familles : **cryptographie** (signatures, trousseaux de clés), **audit** (chaîne vérifiable, aucune feuille sans clair), **fail-closed** (pannes d'OPA, du traducteur, de l'horloge), **quorum** (preuves liées à l'action, à la cellule et à la politique) et **anti-rejeu** (preuves à usage unique, durables après redémarrage). Les correctifs récents sont éprouvés par mutation : on réintroduit la faille et l'on vérifie que le test échoue.

**Conséquence métier.** Les garanties annoncées sont exécutables : un évaluateur lance la commande et voit le résultat. Le contrôle s'exécute à chaque modification du code (intégration continue).

### 4.2 Red team 2026

Une revue adversariale du chemin de décision, menée avec l'assistance de deux systèmes d'IA — Claude (Anthropic) et Cyberkimi — (1<sup>er</sup> octobre 2026 : 5 passes, 25 fichiers, analyse statique du code, sans exploitation dynamique) a produit 22 constats numérotés, dont **un** de sévérité haute (disponibilité : un verrou de sécurité sans route de levée) et **aucune faille de fraude non mitigée**. Le constat de sévérité haute, un constat de sévérité moyenne (corps de requêtes non bornés) et un constat moyen-bas (preuves de quorum rejouables dans leur fenêtre) sont corrigés et couverts par le selftest. Un rapport de vérification du 2 octobre (8 constats) a donné lieu à un correctif fusionné dans le dépôt pour chacun d'eux.

**Conséquence métier.** Les failles trouvées sont tracées une à une et corrigées en public. Une revue statique n'établit pas l'absence de faille : une campagne dynamique reste à mener, le lab de validation comprenant un red team (§6).

### 4.3 Latence et dimensionnement

| Mesure (VM de développement, 4 CPU, 2 oct. 2026) | Résultat                                             | Ce que cela signifie                                                                                |
|--------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Un flux seul, OPA chaud                          | p99 ≈ 2,7 à 3,3 ms (budget 5 ms)                     | Environ 1 ms par décision en régime normal.                                                         |
| Première requête après démarrage à froid         | > 5 ms dans 5 à 33 % des essais                      | Le premier verdict peut être un refus de délai ; un échauffement avant mise en service le supprime. |
| Concurrence croissante                           | Saturation vers 2 à 3 décisions simultanées par cœur | À dimensionner : environ un cœur pour 2 à 3 décisions concurrentes visées.                          |

Ordre de grandeur sur une machine de développement avec une politique d'exemple, pas un résultat de production : le matériel d'un pilote est différent, et le coût dépend directement des règles. Outil et rapports bruts : *tests/opa\_latency* du dépôt TBP-NETWORK.

## 4.4 Vingt et un catalogues de conformité

Chaque cadre de référence a été lu contrôle par contrôle face au code réel. Chaque ligne est **couverte** (mécanisme nommé), **à corriger** (renvoi à une issue) ou **hors du champ de TBP** (avec le geste attendu du déployeur). Ce sont des cartographies, pas des certifications.

| Cadre                                                                                  | Statut            | Lecture                                                                                                                                    |
|----------------------------------------------------------------------------------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| EU AI Act (règlement 2024/1689)                                                        | Complet           | Art. 12 et 14 couverts ; voir §5.                                                                                                          |
| NIST AI RMF 1.0                                                                        | Complet           | TBP est surtout un composant « Manage », avec des apports à « Measure » ; « Govern » et « Map » restent organisationnels.                  |
| ISO/IEC 42001 · 23894 · 27001/27002                                                    | Complet           | Contrôles techniques couverts ; processus de gouvernance hors champ.                                                                       |
| SOC 2                                                                                  | Complet           | Compromis disponibilité / sécurité explicité (fail-closed).                                                                                |
| NIST SP 800-207 (Zero Trust)                                                           | Complet           | Meilleure correspondance de la série : les 7 principes couverts.                                                                           |
| IEC 62443 (OT / industriel)                                                            | Complet           | Isolement et traçabilité ; exigence de disponibilité traitée comme compromis assumé.                                                       |
| MITRE ATLAS · NIST CSF 2.0 / SP 800-53                                                 | Partiel           | Un trou à combler pour les deux : profil comportemental par séquences (issue #181).                                                        |
| OWASP LLM Top 10 · API Security · Agentic Skills                                       | Complet / Partiel | Deux points ouverts sur le LLM Top 10 : profil comportemental (issue #181) et console d'opérateur affichant l'action traduite (issue #86). |
| SLSA · SCVS · CIS · RGPD · FIPS 140-3 · CSA MAESTRO · directives fédérales américaines | Complet           | FIPS 140-3 : cadrage documentaire seulement (pas de validation de module). ISO/IEC 22989 (terminologie) : sans objet.                      |

« Complet » signifie : aucune ligne ne reste à corriger par du code ; les lignes hors champ restent à la charge du déployeur. Source : [tbp-compliance/](#), dépôt TBP-NETWORK.

## 4.5 Ce que TBP ne fait pas

Une barrière qui annonce ce qu'elle ne couvre pas est plus utile qu'une barrière qui promet tout. Les quatre limites suivantes sont structurelles, pas des défauts à corriger.

|                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                  |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>TBP NE JUGE PAS...</b></p> <ul style="list-style-type: none"><li>la <b>qualité du raisonnement</b> du modèle : il juge l'action qui en résulte, pas le chemin pour y arriver ;</li><li>l'<b>exactitude des faits</b> : il ne fait pas de vérification des affirmations du modèle.</li></ul> | <p><b>TBP NE GARANTIT PAS...</b></p> <ul style="list-style-type: none"><li>la protection de ce qui <b>ne passe pas par lui</b> : un agent qui agit par un chemin non gouverné n'est pas contrôlé ;</li><li>la sécurité <b>à lui seul</b> : il s'ajoute à l'isolation des machines, à la gestion des identités et aux autres contrôles.</li></ul> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

**Conséquence métier.** TBP transforme une question insoluble — empêcher toute erreur d'un modèle — en une question que l'on sait traiter : borner ce qu'une erreur peut atteindre, la tracer et savoir qui en répond.

## 5 Conformité EU AI Act

Le règlement ne prescrit pas d'architecture ; il prescrit des démonstrations.

Le règlement (UE) 2024/1689 ne dit pas « installez une couche d'exécution ». Il exige, pour les systèmes à haut risque, une journalisation automatique (art. 12), un contrôle humain effectif (art. 14) et la robustesse face aux attaques (art. 15). Pour un agent qui agit sur des systèmes, ces trois démonstrations ne peuvent se faire qu'à l'endroit où l'action s'exécute : c'est l'argument de ce document, pas une prescription du texte.

| Article               | Obligation                                                                                                           | Statut TBP     | Preuve technique                                                                                                                                                                                                                                                                  |
|-----------------------|----------------------------------------------------------------------------------------------------------------------|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 12                    | Enregistrement automatique, sûr et traçable                                                                          | Couvert        | Chaîne de Merkle, traces « hash-only », signature matérielle ; l'approbation d'un plan est attribuée à la clé de l'opérateur. L'ancrage externe (horodatage RFC 3161) est une brique livrée et testée ; son câblage en production est prévu au premier déploiement multi-cellule. |
| 14                    | Contrôle humain effectif, capacité d'intervenir et de bloquer                                                        | Couvert        | Quorum $k$ -sur- $n$ , co-signature avant action irréversible, pour les agents enregistrés en classe W. Limite : la classe vient du registre des agents, non de l'action (voir §7).                                                                                               |
| 15                    | Robustesse et cybersécurité                                                                                          | Couvert        | Clé HSM non extractible, mTLS, anti-rejeu, fail-closed généralisé, bundle de règles signé. L'exactitude du modèle reste hors champ.                                                                                                                                               |
| 9, 10, 11, 13, 27, 43 | Gestion des risques ; données ; documentation technique ; transparence ; analyse d'impact ; évaluation de conformité | Hors périmètre | Obligations organisationnelles, hors du périmètre technique. TBP fournit des éléments à citer dans ces dossiers (traces, codes de raison stables, cartographies).                                                                                                                 |

« Couvert » : couvert par un mécanisme technique, avec ses limites. « Hors périmètre » : hors du périmètre technique de TBP ; ce n'est pas « non traité », c'est du ressort du fournisseur ou du déployeur. Un statut « couvert » n'est pas une attestation de conformité : celle-ci relève d'une évaluation par l'organisation concernée.

**Conséquence métier.** Pour trois articles, la réponse à « comment le démontrez-vous ? » est un mécanisme vérifiable, pas une procédure. Pour les autres, TBP ne prétend pas répondre à la place de votre organisation.

## 6 Déploiement

Un seul code ; l'échelle est un réglage, pas un produit différent.

| Échelle    | Périmètre                                                                                          | Quorum et supervision                             |
|------------|----------------------------------------------------------------------------------------------------|---------------------------------------------------|
| Scale 1    | Une machine, un administrateur. Point d'application devant le service, OPA local, un registre.     | $k = 1$ . Pas de broker ni de supervision dédiée. |
| Scale 2    | Un petit site derrière un broker unique ; contrôle d'accès réseau (802.1X) recommandé.             | $k = 2$ sur 3 recommandé.                         |
| Scale 3    | Plusieurs cellules (une VM par cellule), bail d'époque, superviseur indépendant en lecture seule.  | $k$ sur $n$ , contrôleurs en HSM.                 |
| Scale full | Plusieurs entités qui se prouvent mutuellement leur politique et la continuité de leur historique. | Non démarré : nouveau travail de protocole.       |

Les guides d'installation des échelles 1 et 2 et du déploiement multi-cellule existent dans le dépôt, avec pour chaque étape un critère de succès vérifiable ; le selftest vérifie leur structure et rejoue leurs scénarios.

### Prérequis

- **Machines.** Une VM par cellule. Dimensionnement : environ un cœur pour 2 à 3 décisions concurrentes visées (§4.3).
- **Clés.** Un HSM pour les clés des contrôleurs (cérémonie de genèse). SoftHSM n'est admis qu'en développement. La v4.2.1 documente des signataires PKCS#11 : YubiKey, AWS CloudHSM, Azure Key Vault. Leur validation sur ces services cloud n'a pas encore été faite.
- **Réseau.** Aucune route de sortie des agents autrement que par TBP (règles de pare-feu et vérification fournies) ; mTLS entre les agents distants et le broker.

**Temps jusqu'au premier verdict** — objectif : 5 minutes. La pile de référence v4.2.1 démarre par *docker compose* (OPA, exemple d'API, métriques) ; le contrôle complet du dépôt TBP-NETWORK s'exécute en une commande (*deploy/selftest/selftest.sh*).

### Le lab de validation (2026)

Un lab de validation est en cours : benchmarks reproductibles de latence et de saturation, red team, dimensionnement, et premiers pilotes chez des opérateurs d'importance vitale (OIV).

## 7 Limites et périmètre

Les limites structurelles figurent au §4.5. Voici l'état de la mise en œuvre au 4 octobre 2026.

### CE QUI EST EN PLACE

- point de décision obligatoire, fail-closed, avec traces signées ;
- quorum  $k$ -sur- $n$  et preuves à usage unique ;
- démarrage mesuré : un fichier de confiance modifié bloque le démarrage ;
- cartographie de 21 cadres ; correctifs de revue tracés une à une.

### CE QUI RESTE OUVERT

- l'ancrage externe n'est pas encore câblé dans un service de production ;
- le chiffrement entre cellules (mTLS inter-cellules) n'est pas implémenté ;
- le quorum d'approbation d'un *plan* n'exige qu'une signature d'opérateur (profil échelle 1 par conception) ;
- aucune certification ni audit externe à ce jour.

Deux autres points à connaître : la classe d'un agent (donc l'exigence de quorum) vient du registre des agents, non de l'action ; un déploiement doit enregistrer en classe W tout agent pouvant accomplir un acte irréversible. Et la saturation de décisions concurrentes se traite par le dimensionnement (§4.3), non par la politique.

**Conséquence métier.** Ces points figurent dans le dépôt, avec leur numéro d'issue et leur état ; un évaluateur peut les vérifier et les suivre.

---

**TBP transforme un problème insoluble — empêcher toute erreur d'un LLM — en un problème solvable : borné, traçable, imputable.**

---

## 8 Références

1. Dépôt TBP v4.2.1, « Shield-Hardening » (Apache 2.0) — [github.com/philippeabraxas-jpg/Responsible-Alliance-Protocol](https://github.com/philippeabraxas-jpg/Responsible-Alliance-Protocol)
2. Dépôt TBP-NETWORK (phase réseau : selftest, guides de déploiement, catalogues de conformité, Apache 2.0) — [github.com/philippeabraxas-jpg/TBP-NETWORK](https://github.com/philippeabraxas-jpg/TBP-NETWORK)
3. Spécification V3.1 (dépôt TBP v4.2.1) et spécification v1.4.10 (dépôt TBP-NETWORK, *docs*)
4. Red team : *Red\_team\_analysis.md* (arguments contre l'adoption) et *docs/audits/redteam01102026* (revue du 1<sup>er</sup> octobre 2026)
5. Mesures de latence : *tests/opa\_latency/README.md* et *results/* (2 octobre 2026)
6. Incident Hugging Face : METR / Redwood Research, août 2026 — [metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation](https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation)
7. Règlement (UE) 2024/1689 sur l'intelligence artificielle (EU AI Act)
8. NIST, *AI Risk Management Framework 1.0* (AI RMF 1.0)
9. ISO/IEC 23894:2023, *Intelligence artificielle — Recommandations sur la gestion des risques*
10. Microsoft, *Agent Governance Toolkit* — [github.com/microsoft/agent-governance-toolkit](https://github.com/microsoft/agent-governance-toolkit) (consulté le 4 octobre 2026)

## Annexe — Termes utilisés

| Terme                | En clair                                                                                  |
|----------------------|-------------------------------------------------------------------------------------------|
| Fail-closed          | En cas de doute ou de panne, la réponse est le refus.                                     |
| HSM                  | Boîtier ou service matériel qui garde une clé de signature sans jamais la laisser sortir. |
| Arbre de Merkle      | Registre où chaque entrée dépend des précédentes : modifier le passé casse la chaîne.     |
| RFC 3161             | Norme d'horodatage par un tiers de confiance.                                             |
| Quorum $k$ -sur- $n$ | $k$ personnes sur $n$ doivent signer avant l'action.                                      |
| OPA / Rego           | Moteur de règles et son langage ; il rend le même verdict pour la même demande.           |
| OIV                  | Opérateur d'importance vitale (infrastructures critiques).                                |

---

Avidost SAS — 130 110 711 RCS Paris. INVARIAN est un produit d'Avidost SAS ; le protocole TBP, conçu par le fondateur d'Avidost SAS, est publié sous licence Apache 2.0. Code source : [github.com/philippeabraxas-jpg/TBP-NETWORK](https://github.com/philippeabraxas-jpg/TBP-NETWORK) (Apache 2.0). Contact : [info@invarian.fr](mailto:info@invarian.fr).