La souveraineté numérique n'est pas un débat politique. C'est une condition de capitalisation de la performance.

Pourquoi le faux dilemme performance versus souveraineté manque la réalité réglementaire de l'IA en environnement régulé, et pourquoi la conformité est une propriété calculable du runtime, pas une déclaration documentaire annuelle.

1. Introduction

La thèse de cet article tient en une phrase : pour une organisation déployant de l'IA dans le périmètre régulé européen, le choix du fournisseur cloud n'est pas un choix de performance arbitré sous contrainte de conformité, c'est un choix de conformité dont la performance est un sous-produit, et dont la stabilité juridique est la condition de capitalisation. Une performance obtenue dans un espace juridiquement instable ne se capitalise pas. Elle s'inscrit comme une dette opérationnelle conditionnelle, exigible au prochain événement réglementaire. Et la conformité elle-même cesse alors d'être une déclaration documentaire annuelle pour devenir une *double propriété d'exécution : calculée par la garde, prouvée par le registre*. C'est l'apport architectural propre de cette doctrine, et le fil qui traverse l'ensemble du texte.

Le calendrier régulatoire en cours transforme la nature de la question pour qui veut déployer en environnement régulé. À compter du 16 mai 2026, tous les certificats HDS actifs en France doivent être conformes à la version 2 du référentiel publié par arrêté du 26 avril 2024. Le 2 août 2026, sauf publication au JOUE du report adopté par le Parlement européen le 26 mars 2026, les obligations applicables aux systèmes d'IA à haut risque listés en annexe III du règlement UE 2024/1689 entrent en application. Le 17 décembre 2025, l'ANSSI a accordé la qualification SecNumCloud 3.2 à S3NS, validant pour la première fois un modèle de souveraineté par le contrôle sur une pile technologique partiellement non européenne. Trois échéances, trois corpus normatifs, un même fond : la conformité de l'infrastructure n'est plus une option d'arbitrage.

Et pourtant, le débat public continue d'osciller entre deux cadrages également stériles, comme si le sujet était encore en phase de discussion philosophique. Le premier oppose protectionnisme et ouverture, comme si le choix d'une infrastructure souveraine relevait d'une préférence idéologique entre Colbert et Smith. Le second oppose performance hyperscaler et cloud souverain dégradé, comme si la conformité n'était qu'une variable

d'ajustement à arbitrer contre le débit IO. Ces deux cadrages partagent le même vice : ils traitent la souveraineté comme un paramètre de configuration, alors qu'elle est une propriété structurelle du système. Et ils confondent invariablement trois notions distinctes que le droit, lui, sépare avec précision : la localisation des données, la qualification de l'infrastructure et la souveraineté juridique de l'opérateur.

Le domaine de validité de cette thèse doit être explicité avant d'aller plus loin, sur deux dimensions distinctes.

D'abord sur le périmètre d'usage. Elle ne vaut pas pour tous les systèmes d'IA, ni pour tous les contextes. Elle vaut pour les systèmes déployés dans le périmètre régulé européen, traitant des données sensibles ou exerçant une fonction à enjeu pour la sécurité, la santé ou les droits fondamentaux des personnes. Hors de ce périmètre, les arbitrages classiques de cloud computing s'appliquent. À l'intérieur, ils ne s'appliquent plus dans leur forme classique.

Ensuite sur la couche traitée. La souveraineté n'est pas monolithique : elle est *stratifiée par couche*, et la souveraineté d'une couche ne supplée pas celle des autres. Au moins quatre couches sont à distinguer : la couche infrastructure cloud (datacenter, IaaS, PaaS, CaaS), la couche modèle (modèles fondationnels, fine-tunes, données d'entraînement), la couche matérielle (accélérateurs GPU, mémoires HBM, fonderies de gravure avancée), et la couche énergétique (alimentation des datacenters dans le mix électrique européen). Le présent article traite principalement de la couche infrastructure cloud, parce que c'est la couche dont les obligations réglementaires sont datées et opposables aujourd'hui (HDS v2, SecNumCloud 3.2, EUCC en cours), et parce que c'est la couche qui s'instrumente architecturalement avec les outils décrits ici. Les couches supérieures sont traitées en discussion comme limites authentiques, et feront l'objet de notes ultérieures dans cette série. Cette restriction est explicite : elle protège l'argument contre l'objection légitime selon laquelle une souveraineté cloud sans souveraineté silicium ni souveraineté énergétique reste une souveraineté logiquement bornée.

2. Clarification terminologique : trois plans, trois menaces

Le mot souveraineté est devenu un agglomérat sémantique qui mélange trois questions distinctes, chacune disposant d'un régime juridique propre, d'instruments d'évaluation propres et de critères d'acceptabilité propres. Confondre ces trois plans, c'est croire qu'on a résolu un problème de conformité parce qu'on a coché une case géographique. C'est, en pratique, l'erreur la plus fréquente dans les comités d'architecture.

Mais la séparation des plans, traitée comme une simple taxonomie, reste descriptive. Pour devenir doctrine, elle doit s'énoncer comme un *modèle de menace* : chaque plan

répond à un type de risque structurellement irréductible aux deux autres. Cette orthogonalité n'est pas terminologique, elle est causale.

Premier plan : la localisation des données. Question géographique : où les bits résident-ils physiquement, et où sont-ils traités. C'est une exigence directe de plusieurs régimes : RGPD (articles 45 et 46), HDS v2, SecNumCloud. *Menace traitée* : le transfert physique et juridictionnel non maîtrisé des données vers un territoire dont le régime de protection est insuffisant ou non équivalent. La localisation neutralise les flux subis. Elle ne neutralise rien d'autre. Un datacenter Azure à Marseille certifié HDS reste opéré par une filiale dont la maison-mère est soumise au CLOUD Act américain. La latitude des serveurs ne neutralise pas la nationalité du gouvernement de la maison-mère.

Deuxième plan : la qualification de l'infrastructure. Question normative : l'hébergeur a-t-il obtenu la certification adéquate pour le type de données concerné. HDS pour la santé, SecNumCloud pour les données sensibles État, OIV et OSE, ISO 27001 pour le socle générique de gestion de la sécurité de l'information, futur EUCC au niveau européen. *Menace traitée* : le défaut de maîtrise opérationnelle et de sécurité de l'environnement d'exécution, indépendamment de la confiance accordée à l'opérateur. La qualification neutralise les défaillances techniques et organisationnelles. Elle ne neutralise rien d'autre. Un hébergeur peut être HDS-qualifié et néanmoins exposé à des injonctions extra-européennes. Pour un établissement de santé en France, héberger des données patient sur un cloud non certifié HDS expose à des risques de non-conformité, de sanctions administratives, contractuelles et éventuellement pénales selon les responsabilités engagées, indépendamment du RGPD. La qualification n'est pas non plus une ISO 27001 rebaptisée : le référentiel HDS v2 et SecNumCloud 3.2 imposent des exigences techniques et organisationnelles qui dépassent largement le socle générique, et qui sont auditées par des organismes accrédités par le COFRAC selon des modalités prescriptives.

Troisième plan : la souveraineté juridique. Question d'extraterritorialité : l'opérateur est-il soumis à des obligations de divulgation extra-européennes, typiquement le CLOUD Act américain de 2018 (Stored Communications Act tel que modifié), la section 702 du FISA, ou les Executive Orders couvrant la collecte SIGINT. *Menace traitée* : la captation extraterritoriale coercitive, c'est-à-dire la capacité d'une autorité étrangère à contraindre l'opérateur à remettre des données ou à modifier ses opérations indépendamment du droit applicable au lieu de stockage. Cette troisième couche est précisément celle que SecNumCloud 3.2 traite frontalement à travers ses critères d'immunité aux lois extracommunautaires, et que le label EUCC en cours de finalisation au niveau européen va structurer pour son niveau « élevé », sur la base des propositions françaises. Une certification HDS française ne neutralise pas un siège social américain.

Le CLOUD Act ne lit pas les datacenters, il lit les organigrammes.

L'orthogonalité des trois plans tient à l'orthogonalité des trois menaces : aucune ne se réduit aux deux autres, aucune n'est annulée par la résolution des deux autres. La localisation est géographique, la qualification est normative, la souveraineté juridique est capitalistique et statutaire. Les trois plans sont nécessaires conjointement pour certaines classes de données, et chacun produit son propre verdict d'admissibilité indépendamment des autres. Un opérateur peut être localisé en France, certifié HDS, et néanmoins exposé au CLOUD Act. Un opérateur peut être immune au CLOUD Act mais ne pas disposer de la qualification HDS. Un opérateur peut être qualifié HDS et immune au CLOUD Act mais ne pas garantir la localisation des opérations de support hors UE. Chacune de ces combinaisons produit un système partiellement conforme, c'est-à-dire, dans le périmètre régulé, non conforme.

Pour le retenir, la triade tient en une phrase. *La localisation dit où sont les bits. La qualification dit ce qu'on a le droit d'y faire. La souveraineté juridique dit qui peut, en dernier ressort, contraindre l'opérateur à les remettre*. Cette structure ternaire est la formulation à laquelle un comité d'architecture doit pouvoir revenir sans ouvrir un texte juridique.

3. Diagnostic : généalogie d'un faux débat

Si la confusion entre ces trois plans persiste à un niveau aussi élevé qu'on l'observe dans les comités d'architecture COMEX et CTO, ce n'est pas par défaut d'information. C'est par effet de cadrage. Trois forces convergent pour entretenir le faux dilemme.

La première force est commerciale. Les hyperscalers américains ont bâti une narration selon laquelle la conformité européenne est une option de configuration, accessible via des « régions souveraines » localisées sur le territoire concerné. Cette narrative est techniquement vraie pour le premier plan (localisation) et pour une partie du deuxième plan (HDS sectoriel). Elle est silencieuse sur le troisième plan (immunité juridique). Le diagnostic doctrinalement exact n'est pas que la narrative serait sélective : la sélectivité est une critique journalistique, pas architecturale. Le diagnostic exact est que *la région souveraine répond à une question différente de celle qu'elle laisse croire résoudre*. La région souveraine répond à une contrainte géographique. Le CLOUD Act agit sur une relation de contrôle capitalistique. Les deux ne vivent pas dans le même espace logique. La région ne peut pas, par définition, neutraliser une menace qui ne s'exerce pas sur la géographie. Ce n'est pas un défaut de communication ; c'est une *incommensurabilité des couches*. Reprocher à la région souveraine de mentir sur le CLOUD Act, c'est reprocher à un altimètre de mal mesurer la latitude.

La deuxième force est culturelle. Le marketing du cloud a passé quinze ans à associer souveraineté et performance dégradée. L'argument typique : le cloud souverain est plus cher, moins outillé, moins agile. Cet argument confond une période historique précise (2018-2022, où l'écart d'outillage était réel et où l'offre française manquait de scalabilité

IA) avec une vérité intemporelle. La qualification SecNumCloud 3.2 de S3NS en décembre 2025, qui couvre simultanément IaaS, PaaS et CaaS et intègre les services data et IA de Google Cloud sous contrôle français, casse l'argument. La qualification de Bleu (Capgemini-Orange-Microsoft) en cours, et l'évolution rapide des offres OVHcloud, Outscale, Scaleway et NumSpot, le confirment : l'écart fonctionnel se réduit. L'argument culturel persiste alors même que sa base empirique s'érode.

La troisième force, plus profonde, est épistémique, et elle tient à un défaut structurel d'organisation : *chaque corpus a son maître ; aucun maître n'a la cohérence*. Le RSSI maîtrise le RGPD, le DSI l'ISO 27001, le CTO le HDS, le directeur juridique le CLOUD Act. Personne, dans la chaîne décisionnelle standard, n'est tenu responsable de l'articulation des quatre corpus comme contraintes simultanées sur une même architecture. Le système est validé par addition de validations partielles, et la cohérence de l'ensemble n'est validée par personne. C'est exactement ce que la doctrine de la *gouvernance-as-architecture*, développée précédemment dans cette série, désignait comme la défaillance structurelle des organisations confondant gouvernance documentaire et gouvernance architecturale.

4. Limites structurelles du paradigme « performance contre souveraineté »

Les limites du paradigme dominant ne sont pas de degré, elles sont de nature. Cinq limites structurelles le rendent inadéquat pour décrire la réalité réglementaire actuelle.

Première limite : l'inversion de causalité. Le paradigme suppose que la conformité est un coût qui réduit la performance. Cette supposition est fausse pour les systèmes en périmètre régulé. Hors qualification, il n'y a pas de système à comparer. La conformité ne réduit pas la performance : elle définit l'espace dans lequel la performance peut être mesurée.

La performance est définie sur l'ensemble des architectures admissibles.

À l'extérieur de cet ensemble, la grandeur « performance » n'a pas de valeur définie, parce que le système n'est pas déployable. Une comparaison entre un cloud hyperscaler non qualifié SecNumCloud et un cloud qualifié SecNumCloud sur des données sensibles État ou OIV n'est pas une comparaison de performance : c'est une comparaison entre un système et l'absence de système.

Deuxième limite, qui prolonge la première : la non-capitalisabilité de la performance non admissible. Même dans les zones où le déploiement temporaire reste possible (par tolérance régulatoire, par défaut d'instruction, par décision politique conjoncturelle), une performance obtenue dans un espace juridiquement instable n'est pas un actif. Elle ne se reconnaît pas au bilan opérationnel comme une capacité durable. Elle s'inscrit, sur le

plan économique, comme une dette conditionnelle exigible à un événement réglementaire futur dont la probabilité d'occurrence n'est pas nulle, voire dont la chronique récente (Schrems I, Schrems II, recours Latombe, instabilité PCLOB-FTC-DPRC après changement d'administration) suggère qu'elle est tendanciellement haute. Le langage technique de l'architecture rejoint ici le langage du contrôle de gestion : la durée d'usage non garantie d'une performance interdit son traitement comptable comme immobilisation. Une architecture qui ne valide pas, par construction, ses conditions d'admissibilité, n'autorise pas l'amortissement de la performance qu'elle produit.

Troisième limite : la confusion entre conformité et certification commerciale. La détention d'une certification ISO 27001, ISO 27701 ou ISO 27018 n'équivaut pas à une qualification SecNumCloud ou HDS, même si les trois corpus partagent une partie de leurs exigences. ISO 27001 est un système de management généraliste auto-déclaratif dans son périmètre, audité contre une norme internationale neutre. SecNumCloud est un référentiel prescriptif spécifique au cloud, qui impose des exigences capitalistiques et juridiques sans équivalent dans la norme ISO. HDS v2 est un référentiel sectoriel sanitaire avec des activités précisément délimitées. Confondre les trois revient à confondre un permis de conduire générique avec une qualification poids-lourd matières dangereuses : les deux attestent d'une compétence, mais ce qu'ils attestent n'est pas commensurable.

Quatrième limite : l'élision du facteur temporel. Le paradigme dominant raisonne en équilibre statique : à un moment donné, l'architecture est-elle conforme. Or les obligations sont datées et évolutives. HDS v2 est en bascule en mai 2026. AI Act annexe III a sa date d'application qui oscille entre août 2026 et décembre 2027 selon l'issue du Digital Omnibus. SecNumCloud 3.2 est en cours de déploiement, avec 12 candidatures en instruction. Le futur EUCC européen pourrait remplacer SecNumCloud à terme.

La conformité n'est pas un état, c'est un flux.

Une architecture conforme aujourd'hui peut ne plus l'être dans 18 mois si elle ne dispose pas d'un mécanisme de tracking réglementaire intégré. Cette propriété temporelle disqualifie les choix architecturaux qui auraient été dimensionnés sur un instantané.

Cinquième limite : l'occultation de la pile algorithmique. Le paradigme dominant raisonne sur la couche infrastructure (IaaS, PaaS, CaaS) et ignore largement la couche logicielle et la couche modèle. Or l'AI Act, pour les systèmes haut risque, impose des obligations directes sur les fournisseurs de modèles et de systèmes (article 16), des obligations sur les déployeurs (article 26), une obligation de marquage CE et d'enregistrement dans la base de données européenne. Ces obligations ne portent pas sur le datacenter, elles portent sur l'objet logiciel et son cycle de vie. La souveraineté de l'infrastructure ne dispense pas de la souveraineté du processus de développement, du contrôle de la chaîne d'approvisionnement des modèles fondationnels, et de la traçabilité des données d'entraînement. Un système IA conforme peut être hébergé sur un cloud souverain et néanmoins reposer sur un modèle fondationnel dont les données

d'entraînement, les biais documentés et la gouvernance de version sont opaques. La conformité de la couche basse ne suffit pas.

À cette cinquième limite il faut ajouter, pour les systèmes IA santé spécifiquement, que la qualification « haut risque » issue de l'annexe III du règlement IA ne résume pas à elle seule le régime applicable. Les systèmes IA santé restent simultanément soumis aux régimes sectoriels MDR (règlement 2017/745) et IVDR (règlement 2017/746), avec leurs mécanismes d'évaluation clinique, de surveillance post-market et de gestion des modifications substantielles. L'articulation des deux régimes est partiellement réglée par le considérant 64 du règlement IA et par les premiers documents de doctrine de la Commission, mais elle reste un objet d'instruction technique pour les organismes notifiés. Une architecture qui prétend à la conformité IA Act sans articulation explicite avec MDR/IVDR n'est pas conforme : elle est partiellement instruite.

5. Architecture : la souveraineté comme propriété structurelle calculable

Si la souveraineté n'est pas un paramètre de configuration, comment se traduit-elle dans une architecture de système IA déployé en environnement régulé. Quatre principes structurent la réponse, organisés autour de deux fonctions distinctes : une fonction de garde et une fonction de registre.

Premier principe : la qualification est un port du système, pas un attribut de l'environnement. Cette formulation reprend la doctrine hexagonale développée dans les articles précédents. Un système est dit souverain non parce qu'il est déployé dans un environnement souverain, mais parce qu'il exige, par construction, un environnement qualifié pour fonctionner. Cette exigence est exposée comme un port externe explicite, vérifié au démarrage, vérifié à chaque opération sensible, et déclenchant un refus structuré (la doctrine WITHHOLD de la *StandardDecisionPolicy*) quand la condition n'est pas remplie. La conséquence pratique : un système d'IA clinique correctement architecturé doit pouvoir refuser de tourner sur un cloud non qualifié HDS, et ce refus est une fonctionnalité positive, pas une dégradation.

Un système qui consent à tourner partout est un système qui ne sait pas où il est légal.

Deuxième principe : la souveraineté est composée, pas achetée. Aucun fournisseur unique ne fournit la pile complète. La pile typique pour un système IA santé en France combine un IaaS qualifié SecNumCloud (S3NS, OVHcloud, Outscale, NumSpot ou équivalent), un orchestrateur conforme (Kubernetes opéré sous contrôle souverain, container CaaS qualifié), un store de données qualifié HDS, des modèles fondationnels dont la chaîne d'approvisionnement est documentée et compatible AI Act (ce qui rend problématique, pour les usages les plus sensibles, le recours à des modèles fermés dont

les données d'entraînement et les fines-tunes ne sont pas auditables), et une couche applicative dont la traçabilité événementielle est assurée par une architecture orientée événements compatible avec les exigences de logging de l'article 12 AI Act. Cette composition n'est pas un produit commercial. Elle est un travail d'architecture, indissociable de l'ingénierie système, et elle est le sujet propre de l'architecte industriel.

Troisième principe : la souveraineté, traitée comme port, devient un état calculable du runtime ; et elle se décompose en *fonction de garde*. Si la souveraineté est exposée comme une condition d'admissibilité testée à l'exécution, alors elle cesse d'être une déclaration documentaire annuelle pour devenir une primitive synchrone : à chaque opération sensible, le système répond à la question « cette opération a-t-elle le droit de se faire ici, maintenant, avec cette identité de workload, sur ce matériel ». La réponse est binaire. La chaîne d'instruments existe déjà dans la pile cloud-native : *policy-as-code* (OPA, Kyverno, Cedar) pour exprimer les règles de manière déclarative et testable, *admission control* Kubernetes pour refuser le déploiement d'un workload sur une node non qualifiée, *workload identity* attestée cryptographiquement et liée à un environnement certifié, *confidential computing* avec attestation à distance (Intel TDX, AMD SEV-SNP, ARM CCA) prouvant à la couche applicative que le matériel et l'hyperviseur sont dans un état non corrompu, *sovereign posture verification* assemblant ces signaux en un attestat continu. Le port de souveraineté hexagonal n'est donc pas un concept architectural abstrait : il a un programme d'implémentation déjà existant dans les piles productives, et il transforme la conformité de propriété déclarative en propriété observable.

Quatrième principe : à la fonction de garde s'ajoute la *fonction de registre*. Là où la garde répond à « peut-on, ici, maintenant », le registre répond à « peut-on, ensuite, prouver ». Les deux fonctions sont distinctes : la garde est synchrone et boolienne, le registre est asynchrone et historique. Un workload peut être correctement bloqué par la garde sans qu'aucune trace utile ne soit produite ; et inversement, une trace exhaustive peut être produite sans qu'aucune décision d'admission ne soit prise. Les deux ports sont nécessaires conjointement, et ils ne peuvent pas être confondus dans une même primitive. L'AI Act exige une journalisation automatique des événements sur toute la durée de vie du système (article 12), une documentation technique conforme à l'annexe IV, et la conservation des logs permettant l'audit et l'explication des décisions. SecNumCloud 3.2 exige des tests d'intrusion tout au long du cycle de vie de la qualification. HDS v2 exige des audits annuels. L'infrastructure de registre est différente de celle de la garde : architecture événementielle avec logs immuables append-only, traçabilité cryptographique des événements clés, stockage qualifié pour les durées de conservation imposées (10 ans pour certains régimes santé), schémas d'événements documentés et stables. C'est ici que l'EDA, abordée dans la note précédente, cesse d'être un choix d'élégance pour devenir la condition technique de la traçabilité régulatoire.

La distinction qui tranche, et qui condense les quatre principes : *un système n'est pas configuré pour être conforme, il est architecturé pour l'être ; et sa conformité n'est pas attestée, elle est calculée pour la garde et tracée pour le registre*.

6. Articulation avec la doctrine antérieure : une homologie structurelle

Cette thèse ne tombe pas du ciel. Elle s'inscrit dans une chaîne d'articles publiés sous le label Twingital Institute depuis plusieurs mois et dont elle constitue le complément réglementaire. La *gouvernance-as-architecture* posait que les exigences de gouvernance ne sont pas des couches documentaires séparables du système mais des contraintes structurelles qui le construisent. *L'architecture événementielle comme complément essentiel de l'IA agentique* posait que la traçabilité, l'asynchronie et la persistance des événements ne sont pas des choix d'optimisation mais des conditions structurelles de la composition agentique. *L'architecture hexagonale appliquée aux systèmes IA cliniques* posait que les ports et adaptateurs de fiabilité (Out-of-Distribution, Calibration, Audit, Refus) sont des composants de premier ordre, pas des wrappers ajoutés a posteriori.

L'homologie n'est pas accidentelle. Le même *pattern* structurel traverse l'ensemble de la série, et il prend la forme générique d'une distinction entre un signal partiel et la propriété système réelle. *Benchmark ≠ production : un score sur jeu de validation n'est pas une garantie de comportement en distribution déployée. Monitoring ≠ gouvernance : un dashboard d'observation n'est pas un dispositif d'arbitrage ni un mécanisme d'imputabilité. Certification ≠ souveraineté : un papier d'audit n'est pas une propriété continue du runtime. Localisation ≠ immunité juridique : une coordonnée géographique n'est pas un statut capitalistique.* À chaque fois, la même erreur : prendre un *proxy* observable pour la propriété qu'il indique seulement partiellement, et confondre la mesure d'un attribut avec sa réalité.

La thèse de souveraineté architecturale prolonge donc la série exactement sur l'axe homologique. Elle traite la conformité réglementaire comme une contrainte structurelle dont la satisfaction est une propriété architecturale du système, pas une option de déploiement. Elle introduit un quatrième port aux ports de fiabilité hexagonaux : un *port de souveraineté* qui vérifie en continu la conformité de l'environnement d'exécution aux exigences réglementaires applicables, et qui déclenche un refus structuré en cas de violation. Ce port n'est pas un gadget réglementaire surimposé : il est de même nature que les ports OOD ou de calibration, parce qu'il porte sur les conditions de validité du système, pas sur sa performance. Et il participe du même programme : *traduire en propriétés architecturales calculables ce que la communauté traite habituellement comme des engagements documentaires*.

Une dernière homologie, déjà signalée en introduction comme délimitation du domaine, mérite d'être reformulée comme prolongement systémique : la souveraineté est stratifiée par couche, et la rigueur architecturale d'une couche n'entraîne pas mécaniquement la rigueur des couches sous-jacentes. La présente doctrine traite la couche infrastructure cloud, parce que des instruments réglementaires opposables y existent aujourd'hui. Elle ne traite ni la couche modèle, ni la couche matérielle (concentration Nvidia/CUDA, dépendance HBM Samsung-SK Hynix-Micron, concentration TSMC sur les nœuds avancés), ni la couche énergétique (alimentation des datacenters dans le mix électrique européen sous contrainte). Une souveraineté cloud rigoureusement architecturée, mais combinée à une dépendance non interrogée des couches matérielles ou énergétiques, reste une souveraineté logiquement bornée. Cette doctrine n'épuise donc pas le sujet ; elle traite la sous-question pour laquelle un cadre réglementaire daté existe et propose une cible architecturale instrumentable. Les autres couches relèvent d'autres outils et d'autres doctrines, et leur traitement appartient à des notes ultérieures de cette série.

7. Insuffisance du paradigme de marché

Le paradigme de marché actuel, dominé par les hyperscalers américains et les offres « Cloud de Confiance » associées (Bleu, S3NS), propose une réponse partielle au problème mais ne le résout pas. La distinction entre les deux modèles maintenant en concurrence sur le marché français mérite d'être explicitée.

Le modèle dit de *souveraineté par le contrôle*, validé par la qualification SecNumCloud 3.2 de S3NS le 17 décembre 2025, repose sur un montage juridique et opérationnel : l'opérateur est une société de droit français entièrement contrôlée par un actionnaire européen (Thales pour S3NS), opère sur une pile technologique sous licence (Google Cloud), avec localisation française des datacenters et des opérations, et garanties contractuelles d'isolation par rapport à la maison-mère technologique américaine. Ce modèle a été reconnu viable par l'ANSSI, qui a ainsi tranché un débat juridique de plusieurs années. Le modèle dit de *souveraineté technologique pure*, défendu par OVHcloud, Outscale, Scaleway et NumSpot, repose sur une pile entièrement européenne : technologie européenne, opérateur européen, capitaux européens. Les deux modèles coexistent désormais et chacun a sa zone de validité.

Pour l'architecte, la conséquence est la suivante : le choix entre les deux modèles n'est pas réductible à une préférence idéologique ou à un calcul de coût direct. Il dépend du type de données et du type d'usage. Pour des données de santé en exploitation routinière, sans agrégation territoriale massive, sous gouvernance documentée, les deux modèles peuvent convenir. Pour des systèmes de médecine prédictive territoriale agrégeant à grande échelle des cohortes nominatives, ou pour des cas d'usage relevant de SecNumCloud avec des contraintes de classification de l'information remontant à l'instruction générale interministérielle 1300 ou équivalent, le modèle de souveraineté

technologique pure offre une garantie supplémentaire qui peut être nécessaire. Le choix doit être motivé par une analyse de risque documentée, pas par une intuition de coût.

L'objection prévisible est l'argument de l'intégration européenne par le futur EUCC : pourquoi se préoccuper de SecNumCloud si une certification européenne unifiée est en cours de finalisation. L'objection ignore deux faits. Le premier : tant que le règlement EUCC n'est pas adopté et n'a pas désigné son niveau « élevé » comme équivalent à SecNumCloud, la qualification française reste l'instrument applicable. Le second : la déclaration conjointe ANSSI-BSI de mars 2026 sur les critères de souveraineté cloud suggère un alignement franco-allemand qui consolide, plutôt qu'il ne remplace, le standard de niveau élevé. Attendre EUCC, c'est attendre une normalisation qui s'établit *par* SecNumCloud, pas *contre*.

8. Instances illustratives

Les instances qui suivent sont des terrains d'implémentation, pas des preuves générales. Elles sont issues de cas d'usage et d'architectures de référence, et n'ont pas vocation à documenter publiquement le statut de certification des systèmes effectivement déployés. Elles montrent à quoi ressemble, en pratique, la traduction architecturale des trois plans de souveraineté pour des systèmes IA santé en environnement européen régulé.

OCTOPUS : étude ambispective de monde réel sur cohorte mNSCLC BRAF V600E, n=184 patients répartis sur cinq pays européens, traités selon des séquences thérapeutiques différenciées. Pipeline de jumeaux numériques entraîné sur 299 features cliniques nominatives extraites de 59 datasets SAS et 37 domaines SDTM, avec génération synthétique conditionnelle, validation TSTR (95,2 % sur les tâches aval, ce qui ne prouve pas l'indistinguabilité statistique générale mais constitue un indice fort de fidélité opérationnelle pour les tâches de simulation), et simulateur contrefactuel SurvTRACE pour l'analyse de séquences thérapeutiques avec événements concurrents. Sur les trois plans : localisation des données dans l'EEE, qualification HDS de l'environnement d'exécution pour les phases nominatives, immunité aux lois extracommunautaires requise pour la phase d'agrégation cross-pays. Le choix de l'opérateur cloud n'a pas été un arbitrage performance, il a été une condition de faisabilité réglementaire.

Sentinelle IA / PREDICARE : programme de médecine prédictive territoriale, agrégation de données hétérogènes (PMSI, SNDS, biologie, exposome) à l'échelle d'un bassin territorial, modèles prédictifs pour le repérage précoce de transitions cliniques. Contrainte structurelle : l'agrégation cross-source à grande échelle est exactement le scénario où la souveraineté juridique devient critique, parce que les requêtes extraterritoriales sur des bases nominatives massives sont précisément ce que SecNumCloud 3.2 traite. La pile retenue ne pouvait pas reposer sur des composants exposés au CLOUD Act, indépendamment de leur certification HDS sectorielle. Le

système est conditionné par cette exigence, et cette conditionnalité est exposée comme un port externe au système.

ToxTwin : système prédictif de toxicité moléculaire, déployé en mode freemium avec quotas et validation calibrée, hébergé sur infrastructure souveraine française : des serveurs avec GPU et VRAM dans nos locaux sous Ubuntu 24.04. Cas plus simple que les deux précédents parce que les données ne sont pas nominatives. Néanmoins, l'AI Act annexe III, dans sa lecture par certains régulateurs, peut classifier des outils prédictifs santé comme systèmes haut risque selon leur usage dans le pipeline de décision clinique, et l'usage en aide à la R&D pharmaceutique soulève la question de l'articulation IVDR pour les modules diagnostiques associés. La traçabilité des décisions, l'enregistrement dans la base de données européenne (article 71), et la documentation annexe IV sont des contraintes architecturales que le système intègre par construction, pas par patch ultérieur.

Ces trois instances ne prouvent pas la généralité de la thèse. Elles montrent qu'à des échelles et des sensibilités différentes, l'architecture concrète obéit aux mêmes principes : la souveraineté est traitée comme un port du système, composée explicitement, et tracée auditablement.

9. Discussion et limites authentiques

Cette doctrine appelle plusieurs limites qui doivent être formulées explicitement, sous peine de perdre la crédibilité de l'argument.

Première limite : le coût d'opportunité de la non-adoption d'innovations rapides. Les hyperscalers américains disposent d'une avance technologique sur certains services IA spécialisés (modèles fondationnels propriétaires, accélérateurs matériels dédiés, écosystèmes d'outils MLOps intégrés). Choisir une pile entièrement souveraine peut conduire à un retard d'accès à certaines capacités de pointe. Cette limite est réelle. Elle se traite par composition : usage de modèles européens ou open-weight compatibles avec les exigences d'audit pour les couches sensibles, et isolation stricte des couches non sensibles si elles utilisent des composants exposés. Elle ne se traite pas par déni.

Deuxième limite : l'incertitude réglementaire sur le calendrier AI Act. L'adoption probable du Digital Omnibus reportant l'application de l'annexe III au 2 décembre 2027 réduit la pression à court terme. Une organisation peut être tentée de différer ses investissements de mise en conformité. Ce calcul est court-termiste : la date est reportée, les exigences de fond ne le sont pas. Construire un système IA en 2026 sans architecture de conformité, c'est planifier un refactoring massif pour 2027. Le coût d'un retrofit dépasse systématiquement le coût d'une architecture initialement conforme.

Troisième limite : la contestabilité résiduelle du Data Privacy Framework. Le DPF a survécu au recours Latombe en septembre 2025, mais les conditions politiques

américaines (modifications de la PCLOB, de la FTC, de la composition de la DPRC) suggèrent que sa stabilité n'est pas acquise. Schrems lui-même a indiqué en février 2026 que la Commission pourrait suspendre l'accord d'elle-même avant tout nouveau recours. Une architecture qui repose sur le DPF pour ses transferts de données vers les États-Unis hérite d'un risque de continuité non négligeable. Cette limite ne renverse pas la thèse, elle la durcit : elle ajoute une raison d'éviter les transferts non maîtrisés vers des juridictions instables.

Quatrième limite : la souveraineté des couches supérieures et sous-jacentes. La pile infrastructure peut être souveraine sans que la pile modèle le soit, et inversement. Un modèle fondationnel européen entraîné sur un cloud non souverain n'est pas plus souverain qu'un modèle américain entraîné sur OVHcloud. La souveraineté des modèles est un sujet en cours de structuration au niveau européen (Mistral, Aleph Alpha, divers consortiums) mais elle n'est pas réglée. Et au-delà du modèle, la couche matérielle (concentration Nvidia, dépendance HBM, fonderies TSMC) et la couche énergétique restent des points de vulnérabilité majeurs que la doctrine présentée ici ne traite pas. Cette limite signale que la couche infrastructure cloud, bien qu'instrumentable réglementairement aujourd'hui, n'est qu'une couche dans une pile dont la souveraineté complète exige d'autres outils, d'autres calendriers et d'autres doctrines.

Cinquième limite : la transition pratique des systèmes existants. Toute organisation a un patrimoine IT et IA existant qui n'a pas été conçu sous ces contraintes. Le passage à une architecture souveraine ne se fait pas par décret. Il exige un plan de transition qui hiérarchise les systèmes selon leur sensibilité, leur exposition réglementaire et leur coût de migration. Cette transition est un sujet d'architecture en soi, qui mérite un article dédié. La présente note pose la cible architecturale, pas la trajectoire.

Sixième limite : le risque de récupération politique. La doctrine de souveraineté architecturale peut être instrumentalisée par des acteurs économiques qui poussent l'argument souverain pour des raisons protectionnistes plutôt que techniques. Cette récupération est un risque pour la crédibilité de l'argument. Le contre-feu est l'exigence de précision : ne pas confondre les trois plans, ne pas substituer la rhétorique politique à l'analyse réglementaire, ne pas valider une offre parce qu'elle est française mais parce qu'elle est qualifiée. La rigueur architecturale est, ici aussi, le meilleur antidote à l'idéologie.

10. Conclusion

Le faux dilemme « performance contre souveraineté » suppose que les deux soient sur le même axe et qu'on puisse arbitrer linéairement entre elles. Ils ne le sont pas. La performance se mesure dans le périmètre des architectures admissibles. Hors de ce périmètre, il n'y a pas d'arbitrage : il n'y a pas de système. Et même là où le déploiement reste temporairement possible par tolérance, la performance obtenue ne se capitalise

pas : elle s'inscrit comme une dette opérationnelle conditionnelle, exigible au prochain événement réglementaire.

Trois plans à séparer, trois plans à valider, trois plans à tracer. *La localisation dit où sont les bits. La qualification dit ce qu'on a le droit d'y faire. La souveraineté juridique dit qui peut, en dernier ressort, contraindre l'opérateur à les remettre.* Aucun des trois plans ne supplée les autres. Aucun n'est une variable d'ajustement. Et chacun répond à une menace que les deux autres ne neutralisent pas : transfert juridictionnel non maîtrisé, défaut de maîtrise opérationnelle, captation extraterritoriale coercitive.

Dans les périmètres soumis à obligation de qualification, un système d'IA clinique déployé sur une infrastructure non qualifiée cesse d'être admissible juridiquement à l'exploitation : sa performance, fût-elle réelle, ne peut pas être capitalisée, et son existence opérationnelle reste suspendue à une tolérance régulatoire dont l'échéance ne dépend pas de l'architecte. La conformité, dans cette perspective, n'est plus une déclaration documentaire annuelle : elle devient un état observable du runtime, calculé au moment de la garde et tracé dans le registre, refusable lorsque les conditions d'admissibilité ne sont plus réunies.

Le faux dilemme étant levé, la question architecturale qui s'ouvre n'est plus celle du choix entre fournisseurs cloud. Elle est celle de la forme exacte que doit prendre le port de souveraineté dans une pile productive : quelles primitives runtime de garde, quelles attestations cryptographiques, quels schémas d'événements pour le registre, quels seuils de refus structuré, quelle articulation explicite avec les régimes sectoriels MDR et IVDR pour les systèmes santé, quelle architecture de transition pour les patrimoines existants, et au-delà, comment traiter les couches modèle, matérielle et énergétique qui restent ici hors champ. Cette question ouverte est le sujet propre de l'architecte industriel, et elle constitue la suite naturelle de cette doctrine. Le débat sur la souveraineté numérique change de nature pour ceux qui veulent déployer. Il ne se clôt pas : il se déplace.