Oteria Cyber School

Mémoire d'alternance, Master 2 cybersécurité

Méthode pour une intelligence artificielle d'entreprise souveraine et résiliente

De la prévention des menaces à la création de valeur

Étudiant	Thomas Housset
Entreprise	Advens
Maître d'apprentissage	Tanguy Ekizian, consultant senior, Advens
Encadrant mémoire	Paul De Montalivet, Oteria
Période d'alternance	Septembre 2024 à août 2026
Année universitaire	2025-2026

Table des matières

1 Mise en contexte **1**

1.1 Advens . . . 1

1.2 L'équipe et mon rôle . . . 1

2 Introduction du projet **2**

2.1 Contexte d'émergence . . . 2

2.2 Objectifs et impact visé . . . 2

2.3 État de l'art . . . 3

2.4 Grandes étapes du projet . . . 4

2.5 Mon rôle dans le projet . . . 4

3 Description du projet **5**

3.1 Le cadre dans son ensemble . . . 5

3.2 Les cinq phases . . . 6

3.3 Focus 1, cartographie des risques des modèles de langage . . . 7

3.4 Focus 2, comparatif d'architectures entre souveraineté et performance . . . 9

3.5 Résilience, tenir quand le fournisseur lâche . . . 10

3.6 Focus 3, cadre de gouvernance opérationnelle . . . 11

3.7 Rattacher un usage à ses obligations réglementaires . . . 12

3.8 Mise à l'épreuve du cadre sur des outils réels . . . 13

4 Exploiter l'IA, la valeur au bout du cadre **15**

4.1 Pourquoi intégrer, et le coût de ne rien décider . . . 15

4.2 L'IA au service de la défense . . . 16

5 Gestion du projet **16**

5.1 Méthodes et outils . . . 16

5.2 Organisation managériale . . . 17

6 Présentation des résultats **17**

6.1 Atteinte des objectifs initiaux . . . 17

6.2 Impacts non prévus . . . 18

6.3 La matrice comparative comme résultat tangible . . . 18

6.4 Premiers enseignements de l'étude complémentaire . . . 18

6.5 Points de comparaison auprès de RSSI . . . 19

7 Analyse critique **19**

7.1 Ce qui a bien fonctionné . . . 19

7.2 Ce qui a moins bien fonctionné . . . 20

7.3 Ce qui aurait pu être fait autrement . . . 20

7.4 L'étude de terrain, une ouverture assumée . . . 20

7.5 Perspective, l'IA agentique déplace le sujet vers l'identité 20
7.6 Analyse de mon apport personnel 21

8 Apprentissage personnel **21**

9 Conclusion **21**

10 Remerciements **22**

Bibliographie **23**

A Annexes **24**
A.1 Glossaire des sigles 24
A.2 Le cadre, document de référence 25
A.3 Modèle de menace complet 26
A.4 Boîte à outils de contrôles techniques 28
A.5 Fiches détaillées des cas d'usage 28
A.6 Exemple déroulé de bout en bout 29
A.7 Grille de maturité, critères de notation 30
A.8 Exemple de grille de maturité 30
A.9 Méthode de notation et détail des scores de la campagne 30
A.10 Arbre de décision 31
A.11 Modèle de coût des architectures 32
A.12 Réversibilité, schéma de la stratégie de sortie 33
A.13 Correspondance du cadre avec les référentiels 34
A.14 Calendrier de l'AI Act 34
A.15 Niveaux de souveraineté 35
A.16 Schéma de flux et frontières de confiance 35
A.17 Politique-as-code, exemple complet 36
A.18 Structure de l'étude complémentaire 37
A.19 Points de comparaison auprès de RSSI, par secteur 37

1 Mise en contexte

Avant de parler du projet, il faut dire d'où je l'ai mené. Mon poste, mon entreprise et mon équipe expliquent en grande partie la forme qu'a prise ce travail et les contraintes que je me suis fixées.

1.1 Advens

Advens est un acteur français indépendant dont la cybersécurité est l'unique métier. L'entreprise réunit plus de six cents experts, en France et dans plusieurs pays européens, et couvre toute la chaîne, du conseil en gouvernance et en conformité à la sécurité offensive, en passant par la réponse à incident et l'intégration de solutions. Son cœur opérationnel est mySOC, son centre de supervision de sécurité managé, une plateforme de détection et de réponse qualifiée PDIS par l'ANSSI.

Cette position a une conséquence directe sur ma façon de travailler. Je n'interviens presque jamais sur le système d'information d'Advens, mais sur celui des clients, avec leurs outils et leur niveau de maturité, qui varie beaucoup d'une mission à l'autre. Selon les sujets, on touche à la gouvernance et au risque, à la sécurisation d'environnements cloud, ou à de la conduite de projet d'intégration. Mon parcours a navigué entre ces registres, et c'est ce qui m'a permis d'aborder l'intégration de l'IA à la fois sous l'angle de la conformité et sous l'angle technique.

1.2 L'équipe et mon rôle

Je travaille comme consultant en cybersécurité, sur un périmètre qui mêle gouvernance, risque et conformité, des sujets techniques, et de la conduite de projet. Concrètement, je suis détaché chez des clients, le plus souvent en autonomie sur un lot précis, parfois au sein d'une équipe plus large pilotée côté client. Cette autonomie compte pour comprendre la suite, c'est elle qui m'a permis de transformer un besoin flou en un livrable concret.

Un consultant détaché doit changer de posture vite, plus conseil face à un responsable sécurité, plus technique face à des ingénieurs, parfois dans la même journée. Et d'une mission à l'autre, la maturité des organisations est très inégale. Une exigence revient toujours, ce qu'on rend doit tenir sans nous une fois la mission finie. Cette contrainte a directement pesé sur ma façon de penser le livrable décrit plus loin, fait pour être repris par d'autres.

Sur ces deux années, mes missions ont couvert plusieurs registres. J'ai mené des évaluations de maturité cyber en fusion-acquisition, sécurisé des environnements cloud Azure, posture, zones d'atterrissage et gestion des identités et des accès à privilèges, fait de l'ingénierie sécurisée en infrastructure définie par le code et sur les chaînes d'intégration continue, conduit des analyses d'écarts NIS2, ISO 27001 et RGPD, et animé des exercices de gestion de crise pour de grands comptes.

Ce mélange explique l'angle que j'ai pris. Une approche purement conformité aurait sans doute donné une politique sur l'IA, une approche purement technique une architecture de référence. J'ai cherché à tenir les deux, parce que c'est ce que les missions demandaient, une réponse qui

parle au RSSI comme à l'ingénieur qui doit brancher l'outil.

Dernier registre, la cheferie de projet d'intégration logicielle. Étude du besoin, comparaison de solutions, intégration, tests, présentation au client, le tout souvent porté seul. C'est cette habitude de mener une intégration de bout en bout qui a servi de matrice au projet décrit dans ce mémoire.

2 Introduction du projet

2.1 Contexte d'émergence

Le projet est né d'un constat répété en mission. Chez les clients, l'IA générative s'installait partout mais presque jamais de façon décidée, un assistant bureautique activé parce qu'il était dans la licence, un modèle branché dans un scénario d'automatisation, une fonction d'IA proposée d'un coup sur un outil où des données personnelles avaient déjà été déposées. À chaque fois, des données partaient vers un modèle tiers sans que personne ait vraiment décidé que cela devait arriver. Et là où la DSI bloquait l'IA, les utilisateurs basculaient sur leurs comptes personnels, le shadow AI, un angle mort que j'ai retrouvé tout au long du projet.

Cette montée des usages s'est traduite dans mes missions. On me sollicitait de plus en plus pour traiter la gouvernance et le risque de l'IA, soit pour cadrer un nouvel usage avant qu'il parte en production, soit pour regarder ce qui tournait déjà et dire si c'était acceptable. Ces deux questions reviennent sans cesse, et elles n'appellent pas la même réponse. La première demande de construire proprement. La seconde demande d'auditer l'existant, de mesurer les écarts et de proposer une remise à niveau. Le projet est parti de là, de l'idée qu'un consultant devait pouvoir répondre aux deux avec un seul cadre.

Les contextes clients étaient variés, et je les présente ici de façon anonymisée, comme je m'y suis engagé. J'ai surtout travaillé avec des organisations sensibles à la résidence de leurs données, un opérateur de services essentiels soumis à NIS2, un grand compte industriel, des acteurs publics qui manipulent des données personnelles. Ce profil n'est pas neutre pour le sujet. Quand une partie de la donnée est soumise à des contraintes fortes, la question de savoir où elle est traitée et sous quel droit cesse d'être théorique.

Il y a enfin une part d'anticipation, construire ce cadre maintenant, c'était préparer Advens et ses clients à un sujet réglementé qui ne fera que grossir.

2.2 Objectifs et impact visé

L'objectif n'était pas de produire une note de plus sur les dangers de l'IA. Il était de livrer quelque chose qu'un consultant ou un ingénieur puisse ouvrir en mission et utiliser le jour même. Un cadre qui dise, pour un usage donné, s'il peut entrer en production, sous quelles conditions, et comment le surveiller ensuite. Un cadre qui serve aussi bien à construire de zéro qu'à évaluer l'existant.

L'impact visé est double. Pour Advens, disposer d'un référentiel commun pour traiter les intégrations d'IA des clients, et d'un support à présenter en interne pour aligner les équipes sur le sujet. Pour les clients, sortir de l'arbitrage binaire entre tout interdire et tout laisser passer, en remplaçant l'intuition par une méthode défendable devant un RSSI.

2.3 État de l'art

Le sujet se tient à la croisée de plusieurs corpus, déjà abondants. L'intérêt n'est pas de les réciter mais de voir ce qu'ils couvrent, où ils se rejoignent, et ce qu'ils laissent à régler à celui qui décide sur le terrain. Le terrain ne manque pas de cadres, il manque d'un pont entre ces cadres et la décision concrète, et c'est là que se situe ce travail.

Sur le plan réglementaire, l'Union européenne a posé l'AI Act, qui classe les systèmes d'IA par niveau de risque et impose des obligations croissantes selon ce niveau [19]. Son application est progressive, et c'est un point que je rappelle aux clients. Les pratiques interdites et l'exigence de littératie s'appliquent depuis février 2025, les obligations sur les modèles à usage général depuis août 2025, puis viennent celles sur les systèmes à haut risque. Ces dernières, initialement prévues pour août 2026, ont été reportées à décembre 2027 pour les usages de l'annexe III [1]. Ce report illustre ce que je répète en mission, on a affaire à une trajectoire mouvante plus qu'à une échéance figée, et poser sa méthode tôt vaut mieux que de courir après chaque date. À côté, la directive NIS2 élargit le périmètre des entités soumises à des obligations de cybersécurité et encadre la notification d'incidents [17], le règlement DORA fait un travail comparable pour le secteur financier, et le RGPD reste la référence dès qu'une donnée personnelle entre dans un traitement d'IA. Ces textes ne parlent pas le même langage, et un usage d'IA concret tombe souvent sous plusieurs d'entre eux à la fois. C'est l'une des difficultés que le cadre a dû absorber.

Du côté des méthodes de gestion du risque, plusieurs références structurent le paysage et couvrent déjà beaucoup. Le cadre du NIST sur le risque de l'IA propose une démarche par fonctions, gouverner, cartographier, mesurer, gérer [13]. La norme ISO/IEC 42001 décrit un système de management de l'IA dans la logique des normes de management déjà connues des entreprises [8]. Le cadre multicouche de l'ENISA organise la sécurité de l'IA en trois niveaux, fondations de cybersécurité, spécificités liées à l'IA, et déclinaison par secteur [3]. Côté industriel, Google a publié son Secure AI Framework, articulé autour de six éléments fondamentaux, des fondations de sécurité jusqu'à la contextualisation des risques [5], et Gartner a popularisé la notion de gestion de la confiance, du risque et de la sécurité de l'IA, qui insiste sur l'inventaire des modèles et l'inspection à l'exécution [4]. Tous donnent une ossature solide. Leur limite commune est de rester à un niveau d'abstraction qui ne dit pas à un ingénieur quoi faire face à un outil précis, au moment de décider.

Pour le contexte français, le guide de l'ANSSI sur les systèmes d'IA générative est la référence la plus proche de mon usage [1]. Il propose trente-cinq recommandations organisées par phases,

1. Digital Omnibus sur l'IA, accord politique des 6 et 7 mai 2026 confirmé par les États membres le 13 mai, provisoire à la date de rédaction. Le détail des reports figure en annexe (figure 8).

entraînement, intégration et déploiement, puis mise en production, et il classe les attaques en trois grands groupes, manipulation, infection et exfiltration. Mon cadre prolonge ce guide, il en reprend le découpage par phases et la logique de contrôles propres à chaque étape, et il descend là où le guide s'arrête, jusqu'à la décision sur un outil précis du marché, avec une entrée par l'audit de l'existant. C'est ce passage à l'acte qui fait son apport.

Pour cartographier les menaces, deux ressources se sont imposées dans la pratique. Le Top 10 de l'OWASP pour les applications à base de grands modèles de langage liste les défaillances les plus fréquentes, de l'injection de prompt à la fuite par les sorties [15]. La base MITRE ATLAS cartographie les techniques d'attaque visant les systèmes d'IA, dans l'esprit de ce que ATT&CK a fait pour la sécurité classique [12]. Ces deux sources sont concrètes, et c'est sur elles que s'appuie le modèle de menace du projet.

Reste la question de la souveraineté, moins normée mais centrale pour mes clients. Un fournisseur d'IA relevant du droit américain expose les données qu'il traite à des mécanismes d'accès extraterritoriaux, au premier rang le Cloud Act, complété par la section 702 du FISA. Utiliser un tel service pour des données personnelles ou stratégiques, c'est accepter ce risque, même quand l'hébergement affiché est européen. Les recommandations des autorités insistent sur la résidence des données et sur la réversibilité, mais elles laissent à chaque organisation le soin de tracer sa ligne. En mission, c'est nous qui aidons le client à la tracer, et c'est l'un des endroits où le cadre ajoute quelque chose aux référentiels existants, qui restent prudents sur ce terrain juridique.

Un dernier pan mérite d'être cité, l'usage de l'IA dans les opérations défensives elles-mêmes, du tri d'alertes à l'aide à l'analyse. Il rappelle que l'IA est aussi un outil pour la sécurité, ce qui rejoint le titre de ce mémoire, passer de la prévention de la menace à la création de valeur.

Au bilan, l'état de l'art fournit des principes solides, des catalogues de menaces et des structures de gouvernance. Ce qu'il laisse plus volontiers de côté, c'est l'aide à la décision concrète, sur un outil donné, avec une donnée donnée, et un client qui attend une réponse. C'est ce dernier pas que le cadre prend en charge, descendre des principes jusqu'à l'arbitrage. J'en donne la correspondance détaillée, phase par phase, avec ces différents référentiels en annexe (tableau 11).

2.4 Grandes étapes du projet

Le projet s'est déroulé en quatre temps. D'abord un travail d'observation et de collecte sur les usages réellement rencontrés en mission. Ensuite la construction du cadre lui-même, sa méthode et ses outils. Puis sa confrontation à des cas d'usage concrets, qui a servi à l'affiner. Enfin sa mise en forme pour un usage en mission et sa présentation en interne. Une étude complémentaire, un questionnaire en ligne resté ouvert, est venue se greffer en parallèle, comme un volet exploratoire dont je discute les limites plus loin.

2.5 Mon rôle dans le projet

J'ai porté ce projet du début à la fin. Personne ne me l'avait commandé, j'ai formulé le besoin moi-même, à partir de ce que je voyais en mission. La méthode, les grilles, le modèle de menace

et la gouvernance, je les ai assemblés et écrits, en m'appuyant sur les référentiels cités plus haut. Les choix structurants sont les miens, raisonner par typologies d'usage plutôt que par produits, faire tenir ensemble conformité et technique, prévoir un mode pour l'existant autant que pour le neuf, et je les rediscute dans la partie critique.

3 Description du projet

C'est le cœur du mémoire. Je décris d'abord le cadre dans son ensemble, puis les trois éléments sur lesquels mon apport a été le plus important, et je termine par leur confrontation à des cas réels.

3.1 Le cadre dans son ensemble

Le livrable central est une méthode d'intégration de l'IA en entreprise, accompagnée de grilles et de modèles directement utilisables en mission comme en interne. C'est un livrable à part entière, conçu pour Advens et au-delà, que ce mémoire présente et analyse, et dont la version de référence est reproduite en annexe. Autrement dit, le mémoire raconte la construction de ce cadre, et le cadre lui-même l'accompagne comme document opérationnel. La figure 1 le résume d'un coup d'œil, comment un usage d'IA, neuf ou déjà en service, devient une décision motivée. J'ai fait deux choix de conception qui le distinguent d'une simple politique.

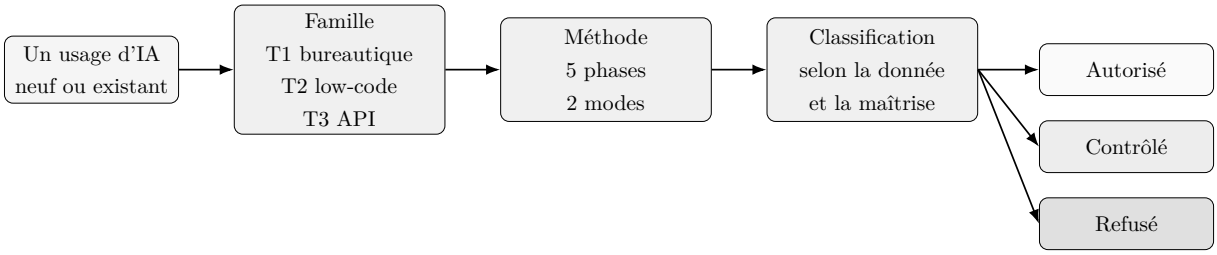

FIGURE 1. Vue d'ensemble du cadre, d'un usage d'IA à une décision motivée, réévaluée quand l'usage ou l'offre évolue.

Le premier choix est de raisonner par typologies d'usage et non par produits. Lister des règles pour tel assistant, puis pour tel autre, puis pour le prochain outil à la mode, c'est courir derrière le marché. J'ai donc regroupé les intégrations en trois familles. Les assistants d'IA dans la suite bureautique, où le modèle accède au périmètre documentaire de l'utilisateur. Les outils low-code ou métier qui embarquent de l'IA, où la donnée circule entre des connecteurs et où l'outil dispose déjà de droits. Les appels d'API d'IA dans des intégrations sur mesure, où tout est à construire mais où l'on contrôle tout. Un nouvel outil se range dans l'une de ces familles, et il hérite de ses points d'attention sans qu'il faille réécrire le cadre. La figure 2 résume ces trois familles et leurs enjeux propres.

Le second choix est de servir deux situations avec une seule méthode. Intégrer une solution de zéro, et évaluer puis améliorer une solution déjà en place. Sur le terrain, c'est souvent la seconde qui se présente d'abord, parce que l'IA arrive avant la gouvernance. La méthode se parcourt alors comme une grille d'audit, chaque point de contrôle devenant un constat tenu ou non tenu, qui

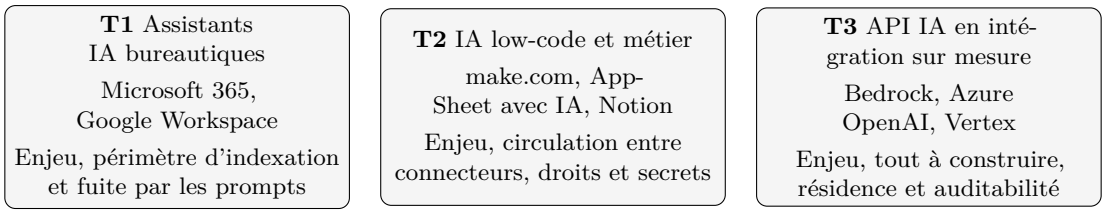

FIGURE 2. Les trois familles d'intégration de l'IA retenues par le cadre.

alimente une grille de maturité et un plan de remédiation. Le même outil sert donc à construire et à corriger, ce qui évite d'entretenir deux référentiels.

La méthode se déroule en cinq phases, chacune fermée par un point de passage. Un point de passage non franchi bloque la mise en production, mais pas le travail d'instruction. Je détaille d'abord ces cinq phases, puis j'approfondis les trois qui portent l'essentiel de la valeur technique et sur lesquelles mon apport a été le plus net.

3.2 Les cinq phases

La figure 3 donne la vue d'ensemble de la méthode, ses cinq phases, les points de passage qui les séparent, et le caractère itératif des allers-retours.

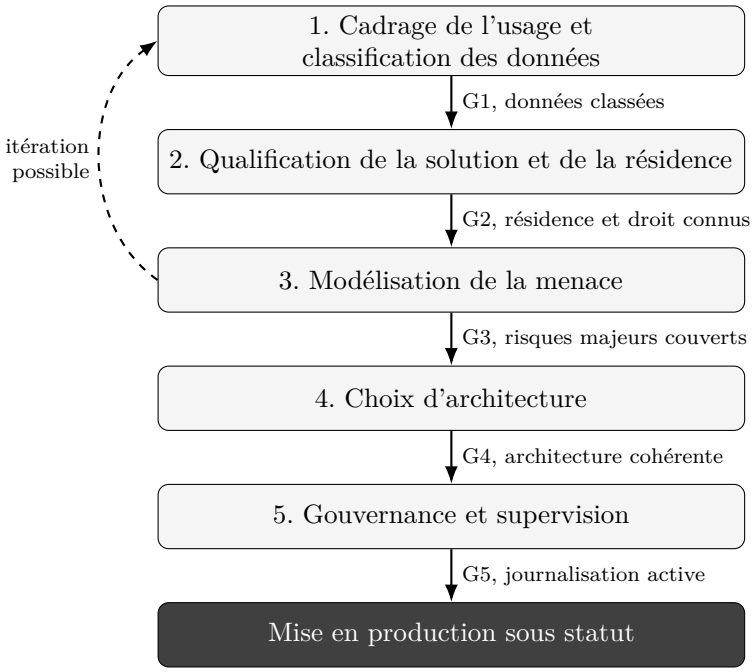

FIGURE 3. La méthode en cinq phases, avec ses points de passage et ses boucles de retour.

La première phase cadre l'usage. On le décrit en une phrase opérationnelle, qui fait quoi, sur quelles données, pour quel gain, et surtout on classe les données concernées, des plus publiques aux données personnelles ou sensibles. Cette étape paraît évidente, elle ne l'est pas. En mission, la première difficulté est souvent que personne n'a posé clairement quelle donnée allait entrer dans l'outil. Le point de passage est simple, sans classification des données, on n'avance pas, parce que c'est elle qui commande tout le reste.

La deuxième phase qualifie la solution et la résidence des données. On établit où vivent et où transitent les données, quels sous-traitants interviennent, s'il y a des transferts hors d'Europe, quel droit s'applique au fournisseur, comment on y accède, et ce que le fournisseur fait des données, les conserve-t-il, les réutilise-t-il pour entraîner ses modèles. Le point dur récurrent se loge ici, un outil qui propose de l'IA sur des espaces où le métier a déjà déposé des données personnelles fait traiter ces données par le fournisseur, et si l'hébergement est aux États-Unis, on accepte de fait l'extraterritorialité. Le point de passage exige une résidence connue, un droit documenté, et un accès au nom de l'entreprise, pas en compte personnel.

La troisième phase modélise la menace, la quatrième choisit l'architecture, la cinquième met sous gouvernance et supervise. Ce sont les trois focus que je développe en détail plus bas, parce que c'est là que se concentre la valeur technique du cadre et que mon apport y a été le plus important. J'y ajoute un volet résilience qui prolonge directement le choix d'architecture, pour que la continuité ne soit pas oubliée une fois la solution retenue. Avant d'y entrer, une remarque sur l'enchaînement. Les phases ne sont pas un tunnel administratif, elles se relancent. Une qualification souveraineté défavorable peut renvoyer au cadrage pour réduire le périmètre de données. Un modèle de menace qui révèle un risque non couvert peut imposer un autre choix d'architecture. C'est ce caractère itératif qui distingue une méthode utile d'une simple liste de cases à cocher.

3.3 Focus 1, cartographie des risques des modèles de langage

La première brique est un modèle de menace de référence, pensé pour être réutilisé sur n'importe quel usage reposant sur un modèle de langage, puis étendu aux applications à IA intégrée.

J'ai construit ce modèle en partant des modes de défaillance documentés, en particulier le Top 10 de l'OWASP pour les applications à base de modèles de langage [15], puis en les rattachant à deux grilles connues des équipes sécurité, STRIDE pour la nature de l'atteinte et MITRE ATLAS pour la technique d'attaque [12]. Chaque risque est ensuite scoré sur trois axes, sa vraisemblance, son impact, et sa détectabilité, sur une échelle simple à trois niveaux. Le troisième axe compte autant que les deux autres. Un risque peu détectable s'installe sans qu'on le voie, et c'est souvent le plus coûteux à long terme. C'est lui qui justifie d'exiger de la journalisation même quand l'impact unitaire paraît faible. Le tableau 1 en donne une version condensée, la version complète, avec le repère ATLAS et le contrôle associé à chaque risque, ainsi qu'une heatmap de priorisation, figurent en annexe.

Le modèle vaut moins par sa liste que par les contrôles que j'y ai associés et la façon dont on le dérive. Pour un usage donné, on ne garde que les risques pertinents, on ajuste le score au contexte, et on tranche, contrôle exigé ou risque accepté de façon tracée.

Un exemple concret de dérivation. Pour un assistant d'IA intégré à la suite bureautique d'entreprise, le risque dominant est la divulgation d'informations par surexposition, devant l'injection de prompt. L'assistant s'appuie sur l'index de recherche de l'organisation et respecte les permissions

Risque	STRIDE	V	I	D
Injection de prompt directe	Tampering	+++	++	++
Injection indirecte via une source	Tampering	++	+++	+
Jailbreak des garde-fous	Elev. of Priv.	+++	++	++
Empoisonnement des données	Tampering	+	+++	+
Exfiltration par les sorties	Info. Disclosure	++	+++	+
Fuite via les prompts des utilisateurs	Info. Disclosure	+++	+++	++
Dépendance au fournisseur	Denial of Service	++	+++	++
Dérive du modèle	-	++	++	+
Dérive de la supervision	Repudiation	++	+++	+

TABLE 1. Modèle de menace de référence, version condensée. V vraisemblance, I impact, D détectabilité. Échelle, plus de signes plus la valeur est élevée, de + faible à +++ élevé.

existantes, et c'est précisément le piège, il rend visible et synthétise d'un coup tout ce qui était déjà mal cloisonné [10]. Une erreur de droits qui dormait inaperçue devient une réponse générée. Le modèle de menace classe ce risque en impact élevé et détectabilité faible. Le contrôle ne se joue donc pas sur le prompt mais en amont, sur l'hygiène des permissions, et il s'outille concrètement, une évaluation de la surexposition avant déploiement par une posture de sécurité des données dédiée à l'IA, des étiquettes de sensibilité qui conditionnent l'accès du modèle au contenu chiffré par un droit d'extraction, et une recherche restreinte à un périmètre validé comme mesure d'attente le temps de corriger les droits. La leçon dépasse un produit, sur ce type d'assistant le vrai sujet de sécurité est la gouvernance des accès existants, pas le modèle lui-même. Le schéma de flux avec frontières de confiance donné en annexe (figure 9) situe les endroits où ces contrôles s'appliquent.

Le risque le plus structurel de ces usages mérite qu'on s'y arrête, l'injection indirecte. Dès qu'un modèle lit des données pour produire sa réponse, une page web, un document, un ticket, un attaquant peut y glisser des instructions, et le modèle ne fait pas la différence entre la donnée et la consigne. Les travaux de Greshake et de ses coauteurs l'ont démontré sur des applications réelles, l'intégration brouille la frontière entre données et instructions, ce qui ouvre des attaques à distance sans contact direct avec l'utilisateur [7]. C'est pourquoi, dans le cadre, le contrôle des sources et le nettoyage du contexte ne sont pas optionnels pour un usage de type RAG, ils sont la contre-mesure de premier rang. L'édition 2025 du Top 10 de l'OWASP a d'ailleurs intégré les risques nés du RAG et des agents, les faiblesses des vecteurs et des plongements, l'agence excessive d'un système qui agit au-delà de ce qu'on attend, la fuite du prompt système, la consommation non bornée qui transforme un usage en déni de service ou en facture incontrôlée. Le modèle de menace reprend ces entrées, l'agence excessive renvoyant à la gouvernance des droits, la consommation non bornée à la supervision des quotas et des coûts.

Pour les applications à IA intégrée, j'ai ajouté les risques propres à ces environnements, des secrets en clair dans un scénario d'automatisation, une surexposition liée à des droits déjà accordés à

l'outil, un périmètre d'indexation trop large. Le reste du modèle continue de s'appliquer, ce qui laisse penser qu'il reste valable au-delà du seul agent conversationnel.

3.4 Focus 2, comparatif d'architectures entre souveraineté et performance

La deuxième brique répond à une question que tous les clients finissent par poser, où faire tourner le modèle. J'ai formalisé trois configurations types et une grille pour les départager, sans désigner de gagnant.

La première configuration repose sur des modèles à poids ouverts, opérés sur un cloud souverain ou en interne. La souveraineté et l'auditabilité y sont fortes, au prix d'un effort d'ingénierie élevé et d'un plafond de capacité plus bas que les très grands modèles. La deuxième configuration appelle un modèle managé d'un grand fournisseur, en région européenne. La capacité et la latence sont au meilleur niveau, la mise en œuvre est rapide, mais le droit applicable au fournisseur reste à surveiller et le risque de dépendance est plus fort. La troisième configuration est hybride, elle route explicitement les traitements sensibles vers le souverain et le reste vers le service managé. On gagne en capacité sans exposer le critique, contre une complexité de routage et de gouvernance.

Critère	Souverain	API UE	Hybride
Souveraineté, résidence	Élevée	Moyenne	Modulable
Droit applicable maîtrisé	Élevé	À surveiller	Élevé sur le sensible
Capacité du modèle	Moyenne	Élevée	Élevée hors sensible
Latence	Variable	Faible	Variable
Coût	Élevé	À l'usage	Intermédiaire
Auditabilité	Élevée	Moyenne	Élevée
Exposition de la donnée	Faible	Élevé	Moyen
Effort d'ingénierie	Élevé	Faible	Élevé

TABLE 2. Comparatif des trois configurations d'architecture.

La figure 4 situe les trois configurations sur les deux axes qui comptent, la maîtrise et la capacité, et montre qu'aucune ne domine l'autre partout. J'y ajoute en repère l'API globale hors d'Europe, celle que le cadre écarte pour des données sensibles. Sa capacité est la même que celle de l'API européenne, ce sont les mêmes modèles, et elle donne parfois accès plus tôt aux toutes dernières versions soumises à des restrictions géographiques, ce qui en fait la tentation, mais elle perd la maîtrise du droit applicable.

Ce tableau 2 sert à rendre l'arbitrage explicite, il ne tranche pas à la place du client. Un RSSI peut s'en saisir pour défendre un choix sur des critères, et non sur une préférence. Pour des données sensibles avec un besoin d'auditabilité, le souverain ou l'hybride s'imposent. Pour un usage sans enjeu de confidentialité où la capacité prime, l'API managée se justifie. Les valeurs fines, coûts et latences, se mesurent par mission, le cadre fournit la structure de la décision. Les ordres de grandeur de coût qui sous-tendent cet arbitrage, prix au million de jetons contre coût d'un GPU souverain, figurent en annexe (tableau 10).

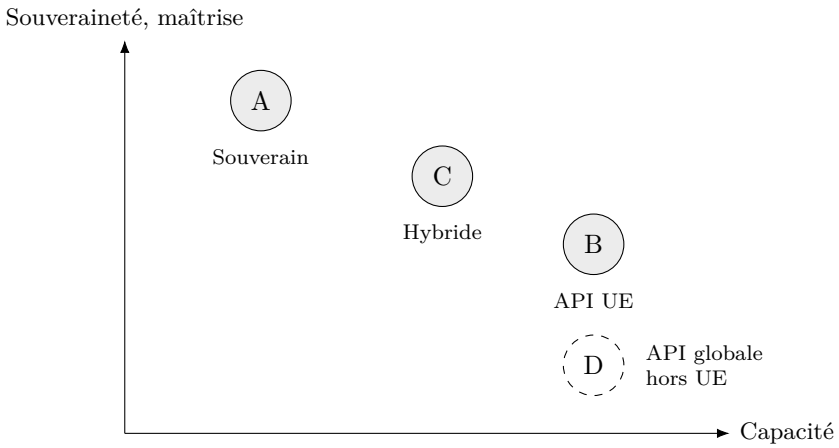

FIGURE 4. Positionnement des configurations. A souverain, B API managée en région UE, C hybride. D, l'API globale hors d'Europe, figure en repère (pointillés), écartée pour des données sensibles.

Un point que le tableau ne montre pas revient à chaque mission, et il mérite d'être posé clairement. Les grands fournisseurs ont nettement progressé sur le plan contractuel. Sur Vertex AI comme sur les offres concurrentes, les données client ne servent pas par défaut à entraîner les modèles, la résidence des données au repos se choisit par des API régionales, et une rétention nulle peut être obtenue par avenant au contrat de traitement pour les clients éligibles [6]. Les API managées en région européenne, Claude sur Amazon Bedrock ou les modèles d'OpenAI sur Azure, offrent des garanties comparables. C'est souvent suffisant pour faire passer un usage en statut contrôlé. Mais cela ne supprime pas le risque de fond, un fournisseur relevant du droit américain reste exposé au Cloud Act quelle que soit la région d'hébergement. La résidence règle le où, pas le sous quel droit. Le cadre force à poser cette distinction, et c'est elle qui sépare un usage acceptable d'un usage à éviter pour des données vraiment sensibles. Les trois niveaux de souveraineté et leurs contrôles compensatoires, dont la pseudonymisation sur site, sont détaillés en annexe (tableau 12).

Un exemple d'arbitrage rendu. Pour un assistant de support traitant des données personnelles de clients, j'écarte l'API globale hors d'Europe, et je compare deux options, le souverain auto-hébergé, plus lourd mais sans exposition juridique, et l'hybride, une API managée en région européenne avec pseudonymisation sur site des champs identifiants. À volume modéré, l'hybride l'emporte, il donne la capacité d'un grand modèle tout en gardant la donnée nominative dans le périmètre maîtrisé. Le même raisonnement, pour des données stratégiques de recherche et développement, bascule vers le souverain, parce que là le risque d'exposition de la donnée ne se compense pas.

3.5 Résilience, tenir quand le fournisseur lâche

La souveraineté répond à la question du droit, la résilience à celle de la continuité. Le titre les associe volontairement, parce qu'un usage d'IA crée une dépendance nouvelle dont peu mesurent la portée. Le risque dépasse la simple panne. Un fournisseur peut changer ses prix, durcir ses conditions, ou retirer le modèle sur lequel on a bâti ses automatisations, et mon étude a justement fait remonter cette dépendance comme le risque le plus sous-évalué.

Le cadre traite ce point par la réversibilité, pensée dès la phase d'architecture. Pour un usage critique, je documente une voie de sortie, quel modèle de repli et quel effort pour y basculer. Migrer d'une API managée vers un modèle à poids ouverts auto-hébergé n'a rien d'instantané, il faut réoutiller la chaîne d'inférence, reprendre les prompts et les garde-fous, puis valider, soit quelques semaines de travail et un coût d'ingénierie réel. Anticiper cette bascule, même sans la déclencher, en fait une option réelle et non un principe affiché. Le schéma de cette voie de sortie figure en annexe (figure 7).

Pour un client du secteur financier, ce raisonnement devient une obligation. Le règlement DORA impose une stratégie de sortie documentée et testée pour les services numériques critiques, ainsi qu'une évaluation du risque de concentration sur un fournisseur unique [18]. La voie de sortie définie en phase d'architecture est précisément la brique technique de cette stratégie de sortie réglementaire, ce qui ancre la résilience du cadre dans un texte contraignant et pas seulement dans une bonne intention.

3.6 Focus 3, cadre de gouvernance opérationnelle

La troisième brique transforme les décisions précédentes en règles tenables dans la durée. Elle s'articule autour d'une grille de classification, d'une politique écrite en format lisible par une machine, et d'un dispositif de supervision.

La grille de classification range chaque usage en trois statuts, autorisé par défaut, contrôlé, interdit. La règle de calibrage tient en une phrase, on monte d'un cran dès que la donnée se sensibilise ou que la maîtrise de la résidence se dégrade. Un usage sur des contenus publics, avec un fournisseur en résidence européenne et sans réutilisation des données, est autorisé. Un usage sur des données internes confidentielles passe en contrôlé, avec des contrôles obligatoires et une revue périodique. Un usage qui enverrait des données sensibles vers un service sans maîtrise de la résidence, ou qu'on ne peut pas journaliser, reste interdit tant que ces conditions tiennent.

Concrètement, un assistant de rédaction interne sur des notes confidentielles, chez un fournisseur en résidence européenne et sans entraînement, se classe en contrôlé, avec filtrage des données sensibles en entrée, classification des sorties et journalisation. Le même assistant, branché en plus sur la base des ressources humaines qui contient des données personnelles, ne change pas d'outil mais change de classe, il faut alors restreindre les sources ou repasser par la qualification. La grille ne juge pas l'outil, elle juge le couple usage et donnée, et c'est ce qui la rend stable quand les outils changent.

Pour que la politique ne reste pas sur le papier, je l'ai écrite sous forme de politique exprimée en code, dans un format versionnable. L'extrait ci-dessous montre l'ossature pour un usage contrôlé.

```
usage: assistant-redaction-interne
classe_donnees: interne_confidentiel
statut: controle
fournisseur:
```

```
  residence: eu
  entrainement_sur_donnees: false
controles:
  entree:
    dlp: true
    prompts_autorises: liste_blanche_redaction
  sortie:
    classification: true
    masquage_dcp: true
  journalisation:
    actif: true
    contenu: [horodatage, utilisateur, hash_prompt, categorie_sortie]
supervision:
  revue: trimestrielle
incident:
  rattachement: nis2_notification
```

La supervision repose sur une journalisation minimale mais réelle, horodatage, utilisateur, empreinte du prompt plutôt que son contenu en clair, catégorie de la sortie et décision des garde-fous. J'y ai adossé un catalogue de schémas d'incidents, chacun avec sa conduite à tenir et son rattachement aux obligations de notification. Pour une entité soumise à NIS2, ces obligations suivent une cadence précise que le catalogue intègre, une alerte précoce sous vingt-quatre heures, une notification circonstanciée sous soixante-douze heures, puis un rapport final dans le mois qui suit cette notification [17]. Rattacher un schéma d'incident propre à l'IA, par exemple une fuite de données personnelles par une sortie de modèle, à cette cadence, c'est ce qui transforme une bonne intention en obligation tenable. La boucle est alors fermée, un usage n'est pas seulement autorisé, il est surveillé et raccroché aux obligations légales du client.

3.7 Rattacher un usage à ses obligations réglementaires

Un même usage tombe souvent sous plusieurs textes à la fois, et le cadre intègre une étape rapide pour le rattacher à ses obligations sans se substituer à une analyse juridique. Je tiens pour cela une table de correspondance, résumée au tableau 3, qui pour chaque cas d'usage coche les textes applicables et rappelle ce que chacun demande concrètement.

C'est un garde-fou, pas un avis juridique, qui évite l'angle mort le plus courant, traiter un usage comme un simple sujet technique alors qu'il déclenche des obligations légales précises. C'est aussi ce qui rend le cadre crédible devant une direction des risques, parce qu'il parle son langage.

Une précision évite un contresens fréquent. Dans la plupart des intégrations, l'entreprise n'est pas le fournisseur du modèle, elle en est le déployeur. Les obligations lourdes du déployeur, supervision humaine, conservation des journaux au moins six mois, information des salariés, ne s'appliquent que si l'usage relève des cas à haut risque de l'annexe III de l'AI Act, par exemple le recrutement

Texte	S'applique quand	Exigence opérationnelle clé
RGPD	Des données personnelles sont traitées	Base légale, minimisation, information des personnes, encadrement de la sous-traitance
AI Act	Le système relève d'un niveau de risque encadré	Transparence, documentation, surveillance humaine selon le niveau
NIS2	L'organisation est une entité essentielle ou importante	Mesures de sécurité, notification d'incident dans les délais
DORA	L'organisation est un acteur financier régulé	Résilience opérationnelle, maîtrise du risque lié aux tiers

TABLE 3. Table de rattachement réglementaire d'un usage d'IA.

ou la gestion d'infrastructures critiques [19], dont l'entrée en application a été repoussée à fin 2027. Pour la grande majorité des assistants de productivité, ce régime ne s'applique pas, et les obligations se réduisent à la transparence de l'article 50 et à la littératie en IA imposée à tous depuis février 2025 par l'article 4. Le cadre intègre cette distinction, il ne déclenche le régime lourd que là où il est dû, ce qui évite une sur-conformité aussi coûteuse qu'inutile.

3.8 Mise à l'épreuve du cadre sur des outils réels

Pour vérifier que le cadre tenait au-delà du papier, je l'ai appliqué aux usages d'IA que j'ai rencontrés en mission. La méthode a été la même pour chacun, instruire les cinq phases comme un audit, noter chaque phase de zéro à trois, relever les écarts, puis définir et, quand c'était possible, mettre en œuvre les contrôles correctifs avant de renoter. C'est l'écart entre les deux notes qui mesure l'effet du cadre. Les usages venaient surtout d'un grand compte qui avait déjà déployé plusieurs briques d'IA, Gemini comme assistant d'entreprise, des applications AppSheet exploitant ses fonctions d'IA (modèles prédictifs et capacités fondées sur Gemini), et de nombreux scénarios make.com en production dont certains à IA intégrée. S'y ajoutaient un usage interne d'un assistant d'IA d'entreprise et l'étude d'un assistant souverain pour un opérateur public. Les organisations ne sont pas identifiables. Chaque note suit la grille de maturité du cadre, appliquée phase par phase à partir de ce qui était réellement en place, et le détail des notations figure en annexe.

Sur l'ensemble, j'ai instruit une douzaine d'usages, que je regroupe en cinq outils représentatifs, les scénarios make.com comptant à eux seuls pour une dizaine d'entre eux. J'ai relevé de l'ordre de vingt-cinq écarts, dont huit jugés prioritaires parce qu'ils touchaient la résidence des données ou l'absence de journalisation. Le tableau 4 synthétise les résultats par outil représentatif.

Quelques constats chiffrés méritent d'être sortis du tableau. Sur make.com, des secrets étaient présents en clair dans plusieurs scénarios sur la dizaine examinée, et certains appels d'IA partaient par des clés étrangères au tenant d'entreprise, deux défauts classiques de ces outils d'automatisation. Sur Notion, les espaces concernés contenaient des données personnelles dont le traitement passait hors d'Europe, ce qui a conduit au seul refus de la campagne sur la couche d'IA. Plusieurs de ces situations méritent d'être racontées en détail.

Outil	Famille	Mat. avant	Mat. après	Risque principal	Statut
Assistant bureautique (interne)	T1	4/15	11/15	Contenus classés à exclure	Contrôlé
Gemini d'entreprise	T1	5/15	11/15	Résidence, API non maîtrisées	Contrôlé
make.com + Vertex	T2	6/15	13/15	API hors tenant, secrets en clair	Contrôlé
AppSheet avec IA	T2	3/15	9/15	Droits et périmètre de données	Contrôlé
Notion IA	T2	4/15	-	Extraterritorialité sur DCP	Refusé en l'état

TABLE 4. Campagne de test du cadre, scores de maturité avant et après remédiation. Détail des notations en annexe.

Le cas le plus représentatif de mon travail est celui des scénarios d'automatisation. Chez un grand compte, des dizaines de workflows make.com tournaient déjà en production, et certains appelaient un modèle d'IA. En les ouvrant un par un, deux problèmes revenaient. D'abord, plusieurs scénarios appelaient l'IA par des clés qui n'étaient pas celles du tenant d'entreprise, des accès personnels ou de test laissés en place, ce qui faisait sortir la donnée du périmètre maîtrisé. Le travail a consisté à tout faire repasser par les API Gemini de l'entreprise via Vertex AI, avec résidence et journalisation. Ensuite, des secrets traînaient en clair dans des étapes de scénario, corrigés par une gestion centralisée. Rien de spectaculaire, mais c'est typiquement la dérive invisible que le cadre est fait pour rattraper.

Un deuxième cas, plus court, vient d'un usage interne. Un assistant d'IA d'entreprise est déployé, mais le cadre interdit de lui soumettre certains contenus, en particulier des éléments classifiés liés à des clients publics ou très sensibles. Autoriser un outil ne revient pas à l'autoriser pour toutes les données, la frontière se trace sur la donnée, et c'est elle qui décide ce qui peut entrer dans le prompt.

Le cas le plus parlant pour la souveraineté est celui de Notion. Le même grand compte l'utilisait déjà largement, avec une gouvernance en place et un risque accepté sur l'outil lui-même. La vraie question n'était donc pas Notion, mais ce que son IA intégrée ajoutait une fois greffée sur des espaces contenant des données personnelles. La qualification a montré une situation moins binaire qu'on ne le croit. Notion ne réutilise pas les données pour entraîner ses modèles et passe par des sous-traitants comme Anthropic et OpenAI avec une rétention nulle en offre entreprise [14]. Le point dur tient à l'hébergement et à la chaîne de sous-traitance, situés aux États-Unis. Activer l'IA sur ces données revenait à accepter une exposition au Cloud Act, indépendamment de toute clause de non-entraînement. Le surcroît de risque apporté par la couche IA n'étant pas couvert, l'usage a été refusé en l'état, en attendant une alternative à résidence maîtrisée. C'est un cas où le cadre sert surtout à dire non, proprement et avec des arguments.

Le dernier cas prend le problème par l'autre bout, la construction plutôt que l'évaluation. Pour un opérateur public de services essentiels, soucieux de ne rien envoyer hors de son périmètre,

j'ai étudié la faisabilité d'un assistant interne en RAG entièrement souverain, bâti sur une offre qualifiée SecNumCloud avec un modèle à poids ouverts. L'intérêt est double, la donnée ne quitte jamais un environnement de confiance et l'auditabilité est complète. La contrepartie est l'effort, il faut opérer l'inférence soi-même, et le RAG rouvre la porte de l'injection indirecte, qu'un contenu piégé dans une source indexée vienne détourner les réponses, ce qui impose un contrôle et un nettoyage des sources. Ce cas illustre la configuration souveraine du comparatif d'architectures en situation réelle, retenue pour la maîtrise et non pour le coût. Les fiches détaillées de l'ensemble des cas figurent en annexe.

4 Exploiter l'IA, la valeur au bout du cadre

Le titre de ce mémoire annonce un passage, de la prévention des menaces à la création de valeur. Cette section quitte un instant le compte rendu de projet pour assumer cette lecture par la valeur, qui est l'angle de mon titre et le sens même du cadre. Un dispositif de sécurité qui ne servirait qu'à interdire raterait ce passage. Toute la logique du mien est inverse, sécuriser pour pouvoir exploiter, faire du contrôle un permis d'avancer plutôt qu'un frein.

4.1 Pourquoi intégrer, et le coût de ne rien décider

La première valeur du cadre est de rendre l'usage possible. Sur le terrain, le métier veut l'IA pour des gains très concrets, rédiger plus vite ou simplement automatiser ce qui est répétitif. Répondre non à cette demande ne la fait pas disparaître, elle se déplace. C'est le mécanisme du shadow AI, observé chez plusieurs clients et confirmé par mon étude, l'interdiction pousse les utilisateurs vers leurs comptes personnels. Chez un client qui avait bloqué les outils d'IA, j'ai vu des opérationnels recopier à la main des extraits de documents internes dans l'application de leur téléphone, puis recoller la réponse dans leur poste de travail, un contournement artisanal mais bien réel qui faisait sortir la donnée par une porte que plus personne ne surveillait. Le coût de ne rien décider est donc double, on perd le gain de productivité et on hérite quand même du risque, en pire, parce qu'il est devenu invisible. Tout l'enjeu du cadre est là, ramener ces usages clandestins dans un périmètre supervisé, là où on peut les voir, les encadrer et les défendre.

Vu sous cet angle, dire oui sous conditions est en soi une création de valeur. Un assistant cadré, avec son périmètre, ses garde-fous et sa journalisation, libère le temps qu'il promet sans ouvrir la porte que l'on craignait. Sur un usage de rédaction et de synthèse documentaire que j'ai accompagné, le temps de première version d'un livrable type s'est réduit de l'ordre d'un tiers, soit quelques heures regagnées par semaine et par utilisateur, un gain modeste à l'échelle d'une personne mais qui compte à celle d'une équipe. Le cadre rend cette performance défendable au lieu de l'opposer à la sécurité. C'est le message que je porte en mission.

Une nuance, la valeur n'est réelle que si le coût reste maîtrisé. Un usage d'IA peut déraper en facture aussi vite qu'il produit du gain, c'est d'ailleurs un des risques du modèle de menace, la consommation non bornée. Le cadre prévoit donc des leviers simples côté gouvernance, un budget de jetons par usage, la mise en cache des prompts récurrents et le traitement par lots,

qui réduisent nettement la dépense. S'y ajoute le choix du modèle le moins cher capable de faire le travail, plutôt que du plus gros par réflexe. La valeur tient autant au gain obtenu qu'au coût qu'on lui laisse prendre.

4.2 L'IA au service de la défense

L'autre versant de la valeur, c'est l'IA mise au service de la cybersécurité elle-même. Les usages mûrissent vite, tri et corrélation d'alertes, aide à l'analyse, reconstitution de documentation, détection fondée sur l'intention plutôt que sur des signatures. Pour une équipe de défense, le gain de temps est réel, une part du tri de premier niveau est absorbée et l'analyse d'un cas s'en trouve raccourcie, ce qui libère les analystes pour les investigations à plus forte valeur.

Mais ce versant porte un paradoxe que le cadre oblige à regarder en face, utiliser un modèle hébergé hors d'Europe pour protéger un système d'information européen. On mobilise alors, comme outil de protection, la technologie même qui crée le risque que l'on veut contenir. La réponse que je retiens tient au contrôle compensatoire plutôt qu'à l'incantation souveraine. La pseudonymisation sur site, par exemple, traite la donnée sensible dans le périmètre maîtrisé avant tout appel au modèle externe, de sorte que le service distant ne voie jamais de donnée identifiante. Il faut rester lucide, cette mesure réduit l'exposition sans l'annuler, et une donnée pseudonymisée reste une donnée personnelle au sens du RGPD. C'est un arbitrage assumé, pas une solution miracle, et c'est précisément la décision que le cadre sert à rendre explicite plutôt qu'à esquiver.

5 Gestion du projet

5.1 Méthodes et outils

Le projet n'a pas été commandé avec un cahier des charges, je l'ai conduit comme une démarche itérative, en partant du terrain. La matière première venait des missions, des situations vues chez les clients, des questions qui revenaient. J'ai d'abord collecté et rangé ces situations, puis j'en ai tiré les typologies et la méthode, et enfin j'ai éprouvé le cadre sur des cas concrets, ce qui m'a obligé à le corriger plusieurs fois.

J'ai traité le cadre comme un produit que l'on maintient, pas comme un document figé. La méthode et ses grilles sont écrites dans un format texte versionnable, et les politiques sont exprimées en code lisible, ce qui permet de suivre les évolutions et de réutiliser les modèles d'une mission à l'autre. Ce choix vient de mon habitude de l'infrastructure définie par le code, je l'ai transposé à un livrable de gouvernance, parce qu'un cadre qui n'est pas versionné vieillit mal et se contredit vite. Le rythme du sujet l'a confirmé, j'ai dû revoir plusieurs fois des éléments au fil de l'eau, une offre qui change, un modèle retiré, une échéance réglementaire repoussée, et le versionnement a permis d'absorber ces révisions sans tout reprendre.

Le rythme a été celui de l'alternance, des avancées par à-coups entre deux missions, avec des points réguliers pour valider les arbitrages structurants. Cette contrainte de temps a eu un effet plutôt sain, elle m'a forcé à viser l'utile.

Concrètement, le cadre s'est construit sur environ huit mois, en marge de mes missions, et je l'ai versionné comme du code, le corps méthodologique en texte structuré, les politiques en YAML, la grille de maturité dans un tableur. Chaque application sur un cas réel produisait une remontée qui corrigeait le cadre, et une douzaine d'usages instruits ont servi de bancs d'essai successifs. Je me suis tenu à une règle simple, ne rien ajouter qui n'ait servi au moins une fois en mission, pour éviter l'inflation théorique. Les outils mobilisés sont volontairement banals, un dépôt versionné, un éditeur de texte, un tableur, parce qu'un cadre qui exige une plateforme dédiée pour être tenu ne survit pas à la première mission pressée.

5.2 Organisation managériale

J'ai porté la conception sans équipe dédiée, en m'appuyant sur les échanges avec mon maître d'apprentissage, des professionnels du métier et plusieurs contextes clients. Mon rôle a donc cumulé plusieurs casquettes, celle qui définit le besoin, celle qui conçoit, et celle qui présente et défend le résultat.

Les interactions se sont jouées sur trois plans. Avec les clients, en mission, où j'appliquais et testais le cadre sur des usages réels sans être maître de tout, le client fixait le périmètre et ce qu'il était possible de faire ou non, et je m'insérais sur le lot relatif à l'IA. Avec Advens, pour faire du cadre un livrable interne, partageable et présentable aux équipes. Avec l'encadrement académique, pour tenir le cap au fil des retours. Cette organisation légère a favorisé la vitesse de décision, et la confrontation des choix s'est faite par les échanges en mission, les présentations internes, des entretiens avec des RSSI français et suisses de plusieurs secteurs, et l'étude complémentaire.

6 Présentation des résultats

6.1 Atteinte des objectifs initiaux

L'objectif premier était de produire un cadre actionnable, utilisable en mission pour les deux situations, construire de zéro et évaluer l'existant. Sur ce point, le résultat est tangible. La méthode existe, ses cinq phases sont outillées, le modèle de menace, le comparatif d'architectures et le cadre de gouvernance sont écrits et ont été appliqués à des cas réels. Le cadre n'est pas resté au stade de l'intention, il a servi à trancher des situations concrètes, et c'est le critère que je retiens en premier.

Sur les indicateurs, je distingue deux choses. J'ai des mesures concrètes, issues de la campagne décrite plus haut, des scores de maturité avant et après remédiation sur des usages réels. Ce que je n'ai pas encore, ce sont des chiffres d'adoption dans la durée, et c'est normal, le cadre est récent et il est fait pour être réutilisé au sein d'Advens sur le temps long. Ce que je peux déjà affirmer, c'est qu'il a conduit à des décisions de classification et à des plans de remédiation concrets, et qu'il a été présenté en interne pour servir de base commune. En ordre de grandeur, la qualification d'un usage, qui se faisait par échanges informels étalés sur plusieurs jours, tient désormais en une demi-journée balisée par la méthode.

Ces résultats de campagne sont ceux d'une première série de tests, pas d'un suivi statistique sur

la durée, mais ils montrent que le cadre produit des décisions et des corrections, pas seulement des constats. Le cas de l'outil dont l'IA a été refusée faute de résidence maîtrisée est, à sa manière, un résultat aussi parlant qu'une remédiation réussie, il montre que le cadre sait aussi dire non quand il le faut.

Par ailleurs, le cadre apporte ses propres instruments de mesure, et c'est l'un des résultats les plus durables. La grille de maturité, qui note chaque phase de zéro à trois, donne un indicateur reproductible pour suivre un usage dans le temps et comparer plusieurs usages entre eux. Le cadre définit aussi les indicateurs à suivre en supervision, taux de blocage des fuites en entrée, nombre d'incidents, volume d'usage. Autrement dit, le projet ne produit pas seulement une méthode, il produit la façon de mesurer si elle fonctionne.

6.2 Impacts non prévus

Deux effets que je n'avais pas anticipés méritent d'être signalés. Le premier est que le mode évaluation de l'existant a pris plus d'importance que prévu. Je pensais surtout outiller l'intégration de nouvelles solutions, et c'est l'audit de ce qui tournait déjà, souvent sans gouvernance, qui s'est révélé le besoin le plus immédiat. Cela a renforcé la dimension audit du projet, et l'a rapproché du cœur de mon métier.

Le second effet est pédagogique. En présentant le cadre, j'ai vu qu'il servait autant à aligner les équipes et à les faire monter en compétence qu'à trancher des cas. Cet effet d'alignement n'était pas un objectif de départ.

6.3 La matrice comparative comme résultat tangible

Parmi les livrables, la matrice de comparaison des architectures, présentée au tableau 2, est le résultat le plus directement réutilisable. Elle transforme une question récurrente et souvent traitée à l'intuition, où faire tourner le modèle, en un arbitrage explicite sur des critères. Un RSSI peut s'en servir pour justifier un choix devant sa direction, ce qui en fait un résultat concret au-delà du seul mémoire.

6.4 Premiers enseignements de l'étude complémentaire

Le cadre s'appuie sur mes missions Advens, un terrain réel mais limité à mes propres clients. Pour élargir le regard, j'ai d'abord multiplié les entretiens individuels, avec des profils variés et des RSSI de plusieurs secteurs, puis j'ai ouvert un questionnaire en ligne au public pour aller plus loin, dont les résultats consolidés seront publiés sur research.thomas.how. Je n'en livre ici que des signaux, pas des résultats fermés. Les premiers répondants sont volontairement contrastés et extérieurs à mes missions, du développeur indépendant à la consultante en protection des données, jusqu'au responsable technique d'une structure de cybersécurité étrangère. L'étude se poursuit, je n'en tire donc aucune statistique, seulement des pistes qui complètent un terrain par nature restreint.

Quelques signaux concordants ressortent malgré son caractère exploratoire, et ils rejoignent les

choix du cadre, ce qui est plutôt encourageant. La première, presque tout le monde a construit en marchant, sans cadre disponible au moment où l'IA générative est devenue un sujet, ce qui confirme que le manque tient moins au risque qu'à l'absence de méthode pour le traiter. La deuxième, la donnée revient comme le vrai point de contrôle, l'un des répondants en fait même son premier principe non négociable, la classification gouverne l'usage, ce qui est exactement le principe directeur du cadre. La troisième, l'interdiction pousse au contournement, le shadow AI par abonnement personnel est cité, et l'activation encadrée est préférée au blocage. La quatrième, la dépendance au fournisseur est jugée sous-évaluée, au-dessus des risques techniques plus visibles. La cinquième, les référentiels orientés menace comme le Top 10 de l'OWASP sont vus comme actionnables, là où les cadres de gouvernance restent perçus comme théoriques pour les petites structures, ce qui rejoint le constat de l'état de l'art.

Je reste prudent sur la portée de ces signaux, l'échantillon est encore réduit. Mais voir des praticiens de profils éloignés retomber sur les mêmes principes que ceux dégagés en mission est une forme de validation externe que je n'attendais pas si nette.

6.5 Points de comparaison auprès de RSSI

Au-delà de mes missions et du questionnaire, j'ai échangé avec des RSSI français et suisses des secteurs de l'énergie, de la finance et de l'industrie, sur ce qu'ils avaient déjà en place et sur leur façon de traiter le risque. Ces entretiens m'ont donné des points de comparaison que mon seul terrain ne pouvait pas fournir. La finance ressortait comme la plus avancée sur le risque fournisseur et les stratégies de sortie, poussée par DORA côté français et par les attentes de la FINMA côté suisse. L'énergie, en tant qu'opérateur de services essentiels, partait d'une posture plus restrictive et plus sensible à la souveraineté. L'industrie, davantage tirée par la productivité, vivait surtout le shadow AI et la protection de son patrimoine. Côté suisse, la résidence des données sur le territoire et la nLPD revenaient comme une exigence de premier rang, distincte du cadre européen mais convergente sur le fond. Ces différences confirment que le curseur entre souveraineté et usage se règle selon le secteur et le pays, ce que le cadre prend en charge par sa souveraineté graduée. Le détail anonymisé figure en annexe (tableau 13).

7 Analyse critique

C'est la partie où je dois être le plus honnête, y compris avec moi-même. Un cadre porté par une personne, en parallèle de missions, a forcément des forces et des angles morts, et les nommer vaut mieux que les masquer.

7.1 Ce qui a bien fonctionné

Le choix de raisonner par typologies d'usage plutôt que par produits a tenu ses promesses. Chaque fois qu'un nouvel outil apparaissait en mission, il se rangeait dans une des trois familles sans que j'aie à réécrire le cadre. C'est la décision qui a le mieux tenu dans le temps, elle protège le cadre de l'obsolescence.

Le double mode a lui aussi bien tenu. Intégrer une solution neuve ou auditer une solution déjà

en place avec la même méthode a évité de disperser l'effort, et c'est l'audit de l'existant qui a le plus servi. Enfin, écrire la gouvernance en code a rendu le livrable réutilisable et versionnable, ce qui change beaucoup à l'usage.

7.2 Ce qui a moins bien fonctionné

Le premier point faible, c'est moi. Même en confrontant mes choix à mon maître d'apprentissage, à des professionnels, à des RSSI de plusieurs secteurs et pays et à des contextes clients variés, un cadre porté par une seule personne hérite de ses habitudes et de ce qu'elle connaît le mieux. Mon parcours mêle gouvernance et technique, mais je suis sans doute plus à l'aise sur la qualification du risque que sur le volet économique des architectures, et le comparatif s'en ressent, il structure l'arbitrage sans prétendre le trancher en chiffres.

Le second point faible tient à ce qu'une partie des contrôles proposés reste théorique tant qu'ils n'ont pas été déployés à grande échelle. J'ai validé la méthode sur des cas réels, mais certains éléments, en particulier le catalogue d'incidents et son rattachement aux obligations de notification, demandent à être éprouvés dans la durée pour être pleinement crédibles.

7.3 Ce qui aurait pu être fait autrement

Avec le recul, j'aurais pu m'adosser plus tôt et plus explicitement aux cadres existants comme celui de l'ANSSI ou de l'ENISA, ce qui m'aurait fait gagner du temps de structuration. Retrouver chez eux les mêmes grands principes m'a en réalité conforté dans ma direction, et n'enlève rien à ce que j'ai ajouté par-dessus, l'application concrète sur des outils du marché, le double mode audit et intégration, l'arbitrage de souveraineté chiffré et la gouvernance en code, que ces référentiels ne descendent pas jusqu'à fournir.

7.4 L'étude de terrain, une ouverture assumée

L'étude de terrain n'est pas une faiblesse du travail, c'est une ouverture que je revendique. Je l'ai menée pour regarder au-delà de mes seuls clients, d'abord par des entretiens individuels avec des praticiens et des RSSI de plusieurs secteurs, puis par un questionnaire ouvert au public pour aller plus loin. Elle reste exploratoire et se poursuit, je n'en tire donc pas de conclusions générales, mais des recoupements qui éclairent le cadre. La présenter pour ce qu'elle est, un complément vivant plutôt qu'une preuve fermée, me paraît plus juste que de lui faire dire plus qu'elle ne peut.

7.5 Perspective, l'IA agentique déplace le sujet vers l'identité

Le cadre a été pensé pour des assistants et des intégrations où un humain reste dans la boucle. La vague qui arrive, les agents autonomes qui enchaînent des actions et appellent des outils via des protocoles comme MCP, déplace le problème. Quand un agent décide seul d'invoquer un outil à partir d'une description en langage naturel, le brouillage entre donnée et instruction évoqué plus haut devient une surface d'attaque permanente, et le vrai point de contrôle n'est plus le prompt mais l'identité, ce que l'agent a le droit de faire et au nom de qui. Les signaux sont déjà là, la multiplication des identités non humaines et de leurs secrets mal gérés est devenue un sujet majeur en 2026, et les travaux de recherche sur la sécurité des systèmes agentiques se multiplient

[9]. Mon cadre n'épuise pas cette question, mais ses cinq phases y restent utiles, à condition d'ajouter une qualification des droits de l'agent dès le cadrage et un contrôle d'accès fondé sur l'intention en gouvernance. C'est la direction que je donnerais à une version suivante.

7.6 Analyse de mon apport personnel

Mon apport tient au passage à l'acte, transformer des principes déjà posés par les référentiels en une méthode qu'un consultant ouvre et utilise vraiment, sur des outils du marché, dans des missions où la donnée est sensible et la souveraineté un enjeu réel. C'est précisément ce que ces référentiels ne descendent pas jusqu'à faire. Mes axes de progrès, je les connais aussi. Mieux mesurer dans la durée, d'abord. Confronter mes choix à d'autres regards encore plus tôt, ensuite. Creuser enfin l'économie des architectures, que je n'ai fait qu'effleurer. Je les vois comme des chantiers ouverts.

8 Apprentissage personnel

Ce projet m'a appris des choses que les missions seules ne m'avaient pas encore montrées. La première est qu'un bon livrable de sécurité ne vaut que s'il est utilisable. J'avais tendance à soigner la rigueur d'une analyse, j'ai compris qu'une grille qu'on ouvre vraiment en réunion vaut mieux qu'un rapport parfait que personne ne lit. Concevoir pour l'usage, pas pour l'exhaustivité, c'est un réflexe que je garderai.

J'ai aussi mesuré ce que vaut le fait de s'appuyer sur l'existant. En commençant, j'avais l'envie un peu naïve de construire ma propre méthode. En lisant les cadres de l'ANSSI, de l'ENISA ou du NIST, j'ai vu qu'une grande partie du chemin était déjà tracée, et que la maturité consiste autant à savoir réutiliser et adapter qu'à inventer. C'est moins gratifiant sur le moment, mais plus solide à l'arrivée, et c'est finalement ce que j'ai fait, en adossant ma méthode à ces cadres plutôt qu'en cherchant à les refaire.

Sur le plan professionnel, avoir porté ce projet du besoin jusqu'à la présentation m'a fait toucher du doigt mes points forts et mes limites. Je sais formuler un besoin que personne ne m'a dicté et le mener jusqu'à la livraison. Je sais aussi, maintenant, qu'il faut provoquer la contradiction plus tôt, parce qu'à trop avancer seul on finit par s'enfermer dans ses évidences.

Enfin, le sujet lui-même m'a appris à tenir deux exigences qui semblent s'opposer, sécuriser et permettre. Le réflexe d'un débutant en cybersécurité est de dire non. Ce projet m'a convaincu que la vraie valeur est ailleurs, dans la capacité à autoriser sous conditions vérifiables, à rendre un usage possible plutôt qu'à le bloquer. C'est la leçon la plus utile pour la suite de mon métier.

9 Conclusion

Le projet est parti d'un décalage simple, l'intelligence artificielle entrait dans les entreprises plus vite que la gouvernance qui aurait dû l'accompagner. La réponse que j'ai construite chez Advens ne prétend pas régler le sujet, elle propose une méthode pour le tenir, une démarche en cinq phases qui sert aussi bien à intégrer une solution de zéro qu'à évaluer et corriger ce qui

tourne déjà, outillée par un modèle de menace, un comparatif d'architectures, une exigence de réversibilité et un cadre de gouvernance exprimé en code.

Si je devais retenir une idée, ce serait celle du titre. Sécuriser l'IA n'a de sens que si cela permet de l'exploiter. Interdire pour de bon ne tient pas, les utilisateurs contournent, je l'ai constaté en mission. Tout autoriser ne tient pas davantage. Reste une voie étroite, ouvrir l'usage en l'encadrant et rendre cet encadrement vérifiable. C'est là que la souveraineté, la maîtrise des données et la réversibilité jouent leur rôle, celui de garde-fous qui rendent un usage défendable.

Je reste mesuré sur sa portée. Le cadre s'inscrit dans le sillage de références plus établies, et il demande encore à être éprouvé dans la durée. Mais il existe, et il a déjà servi à trancher des cas réels. La suite, je l'ai esquissée dans l'analyse critique, il reste surtout à le confronter à d'autres mains que les miennes.

Une dernière chose, et c'est sans doute la plus importante. Le sujet évolue si vite qu'un cadre figé serait vite dépassé, ces huit mois l'ont confirmé. Le mien est fait pour vivre et être tenu à jour. Il sert déjà en mission, et je le maintiens sur mon site de recherche, research.thomas.how, avec ce mémoire, le modèle pratique du cadre et le questionnaire encore ouvert. Mon objectif est qu'il devienne une référence partagée, réutilisée d'une mission à l'autre chez Advens, et qu'il continue d'évoluer avec les usages et la réglementation. C'est dans cette réutilisation, autant que dans ces pages, que se mesurera sa valeur.

10 Remerciements

Je remercie Tanguy Ekizian, mon maître d'apprentissage chez Advens, pour son accompagnement et pour la confiance qui m'a laissé porter ce projet en autonomie. Je remercie Paul De Montalivet, mon encadrant à Oteria, pour ses retours qui ont permis de bien cadrer ce travail.

Je remercie les équipes d'Advens avec qui j'ai échangé sur ce sujet, ainsi que tous les professionnels rencontrés au fil de ces mois, celles et ceux qui m'ont orienté vers des recherches ou des technologies, et toutes les personnes qui ont pris le temps de répondre à mon enquête de terrain sur mon site de recherche. Je remercie enfin mes proches pour leur soutien pendant toutes ces années d'alternance.

Références

[1] ANSSI. Recommandations de sécurité pour un système d'IA générative (ANSSI-PA-102). Agence nationale de la sécurité des systèmes d'information, 4 2024.

[2] Anthropic. Claude API, tarification par million de jetons. anthropic.com, consulté en juin 2026, 2026.

[3] ENISA. Multilayer Framework for Good Cybersecurity Practices for AI. European Union Agency for Cybersecurity, 6 2023.

[4] Gartner. AI Trust, Risk and Security Management (AI TRiSM). Gartner, 2024.

[5] Google. Secure AI Framework (SAIF), a conceptual framework for secure AI systems. Google, 2023.

[6] Google Cloud. Generative AI, Privacy, and Google Cloud, gouvernance des données Vertex AI. Google Cloud, 2024.

[7] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For : Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. arXiv :2302.12173.

[8] ISO/IEC. ISO/IEC 42001 :2023, Information technology, Artificial intelligence, Management system, 2023.

[9] Sahaya Jestus Lazer, Kshitiz Aryal, Maanak Gupta, and Elisa Bertino. A Survey of Agentic AI and Cybersecurity : Challenges, Opportunities and Use-case Prototypes. arXiv :2601.05293, 1 2026.

[10] Microsoft. Microsoft 365 Copilot, Restricted SharePoint Search et permissions Microsoft Graph. Microsoft Learn, 2025.

[11] Mistral AI. Mistral AI, tarification des modèles. mistral.ai, consulté en juin 2026, 2026.

[12] MITRE. MITRE ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE Corporation, 2024.

[13] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST, 2023.

[14] Notion. Notion AI security and privacy practices. Notion Help Center, 2025.

[15] OWASP Foundation. OWASP Top 10 for Large Language Model Applications, édition 2025, 2025.

[16] Scaleway. Scaleway, tarification des instances GPU H100. scaleway.com, consulté en juin 2026, 2026.

[17] Union européenne. Directive (UE) 2022/2555 (NIS2) concernant des mesures destinées à assurer un niveau élevé commun de cybersécurité dans l'Union. Journal officiel de l'Union européenne, 2022.

[18] Union européenne. Règlement (UE) 2022/2554 sur la résilience opérationnelle numérique du secteur financier (DORA). Journal officiel de l'Union européenne, 2022.

[19] Union européenne. Règlement (UE) 2024/1689 du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (AI Act). Journal officiel de l'Union européenne, 2024.

A Annexes

Les annexes regroupent les versions détaillées des livrables, pour ne pas alourdir le corps du mémoire.

A.1 Glossaire des sigles

Sigle	Signification
IA	Intelligence artificielle
LLM	Grand modèle de langage (Large Language Model)
DCP	Donnée à caractère personnel
RGPD	Règlement général sur la protection des données
AI Act	Règlement européen sur l'intelligence artificielle
NIS2	Directive européenne sur la sécurité des réseaux et des systèmes d'information, deuxième version
DORA	Règlement sur la résilience opérationnelle numérique du secteur financier
RAG	Génération augmentée par la recherche (Retrieval-Augmented Generation)
STRIDE	Modèle de catégorisation des menaces (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege)
MITRE ATLAS	Base de techniques d'attaque visant les systèmes d'IA
NIST AI RMF	Cadre de gestion du risque de l'IA du NIST
ENISA	Agence de l'Union européenne pour la cybersécurité
ANSSI	Agence nationale de la sécurité des systèmes d'information
RSSI	Responsable de la sécurité des systèmes d'information
DLP	Prévention de la fuite de données (Data Loss Prevention)
API	Interface de programmation applicative
ESN	Entreprise de services du numérique
OWASP	Open Worldwide Application Security Project, dont le Top 10 des risques des applications à base de LLM
PDIS	Prestataire de détection d'incidents de sécurité, qualification de l'ANSSI
SecNumCloud	Visa de sécurité de l'ANSSI pour l'informatique en nuage de confiance

Sigle	Signification
Cloud Act	Loi américaine d'accès extraterritorial aux données (Clarifying Lawful Overseas Use of Data Act)
FISA	Loi américaine de surveillance du renseignement étranger, dont la section 702
nLPD	Nouvelle loi fédérale suisse sur la protection des données
FINMA	Autorité fédérale suisse de surveillance des marchés financiers
MCP	Protocole de connexion d'outils pour agents (Model Context Protocol)

A.2 Le cadre, document de référence

Cette annexe reproduit l'ossature du livrable tel qu'il est utilisé en mission, indépendamment du récit du mémoire.

Principes directeurs. La donnée commande, c'est sa classe qui fixe le niveau d'exigence, pas la marque de l'outil. La souveraineté est graduée, un curseur entre résidence, droit applicable et réversibilité. Le compte entreprise est la règle, jamais un compte personnel. Le contrôle est proportionné au risque. Enfin, ce qui ne peut pas être journalisé ne va pas en production.

Trois familles d'intégration. T1, assistants dans la suite bureautique, où l'enjeu est le périmètre d'indexation et la surexposition des droits existants. T2, IA embarquée dans des outils low-code ou métier, où l'enjeu est la circulation entre connecteurs, les secrets et les clés hors tenant. T3, API appelées dans des intégrations sur mesure, où l'enjeu est le contrôle de bout en bout, la résidence et l'auditabilité. Un nouvel outil hérite des points d'attention de sa famille.

La méthode en cinq phases. Le tableau 5 en donne les entrées, le point de passage et le livrable.

Phase	Point de passage	Livrable
1. Cadrage de l'usage	Données classées, porteur identifié	Fiche d'usage
2. Qualification souveraineté	Résidence et droit connus, accès entreprise	Note de qualification
3. Modélisation de la menace	Risques majeurs couverts ou acceptés et tracés	Modèle de menace de l'usage
4. Choix d'architecture	Architecture cohérente avec la classe de données	Décision d'architecture argumentée
5. Gouvernance et supervision	Usage classé, politique active, journalisation	Fiche de gouvernance et politique

TABLE 5. Les cinq phases du cadre, leurs points de passage et leurs livrables.

Qui fait quoi. R réalise, A approuve, C consulté, I informé. Le consultant cyber instruit et propose, le RSSI tranche et porte le risque, le métier reste responsable de l'usage et de sa donnée, le DPO intervient dès qu'il y a des données personnelles, et l'IT outille la journalisation.

Phase	Métier	Consultant	RSSI	DPO	IT
1. Cadrage	A	R	I	C	I
2. Qualification souveraineté	I	R	A	C	C
3. Modélisation de la menace	I	R	A	I	C
4. Architecture	I	R	A	I	C
5. Gouvernance et supervision	I	R	A	C	R

TABLE 6. Répartition des rôles par phase.

Classification des usages. Autorisé par défaut, pour des données publiques ou internes non sensibles chez un fournisseur en résidence européenne sans réutilisation. Contrôlé, pour des données confidentielles ou des données personnelles limitées, sous contrôles obligatoires et revue. Interdit, pour des données sensibles sans maîtrise de la résidence ni du droit, ou tout usage non journalisable, tant que ces conditions tiennent. On monte d'un cran dès que la donnée se sensibilise ou que la maîtrise de la résidence se dégrade.

Deux modes d'emploi. Le cadre se parcourt soit en intégration, avant la mise en production d'un usage neuf, soit en évaluation, comme une grille d'audit appliquée à un usage déjà en service, qui alimente alors la grille de maturité et un backlog de remédiation.

A.3 Modèle de menace complet

Version étendue du tableau 1, complétée des risques propres au RAG et aux agents, avec le repère MITRE ATLAS et le contrôle type associé à chaque risque.

Risque	Repère ATLAS	Contrôle type
Injection de prompt directe	LLM Prompt Injection	Filtrage entrée, séparation instruction et donnée, garde-fous
Injection indirecte via une source	LLM Prompt Injection (indirecte)	Contrôle des sources, signature, nettoyage du contexte
Jailbreak des garde-fous	LLM Jailbreak	Garde-fous en sortie, tests adverses récurrents
Empoisonnement des données	Poison Training Data	Maîtrise des jeux de données, provenance
Exfiltration par les sorties	LLM Data Leakage	Classification des sorties, masquage, revue
Fuite via les prompts	LLM Data Leakage	Sensibilisation, prévention de fuite, pas de données sensibles en prompt
Dépendance au fournisseur	-	Réversibilité, architecture hybride, sortie de secours
Dérive du modèle	Erode ML Model Integrity	Supervision qualité, tests de non-régression
Dérive de la supervision	-	Journalisation, revue périodique, indicateurs
Agence excessive (agent)	LLM Plugin Compromise	Moindre privilège, validation des actions sensibles, identité dédiée
Faiblesses vecteurs et plongements	RAG poisoning	Contrôle d'accès aux index, cloisonnement par tenant, validation des sources
Fuite du prompt système	System Prompt Extraction	Aucun secret dans le prompt système, séparation des secrets
Consommation non bornée	Denial of ML Service	Quotas, limites de débit, surveillance des coûts

Table 7. Modèle de menace de référence, version complète.

La figure 5 en donne une lecture de priorisation, en croisant vraisemblance et impact. Elle montre que la plupart des risques se concentrent du côté du fort impact, et que la case la plus critique, vraisemblance et impact élevés, est la fuite de données par les prompts.

FIGURE 5. Heatmap de priorisation des risques, vraisemblance contre impact. Les numéros suivent l'ordre des lignes du tableau 1.

A.4 Boîte à outils de contrôles techniques

Le modèle de menace dit quoi traiter, cette boîte à outils dit avec quoi. À chaque risque dominant correspondent des contrôles concrets, choisis selon la typologie et la classe de données.

- Injection directe et jailbreak, séparation stricte de l'instruction et de la donnée, marquage des entrées (spotlighting, datamarking), hiérarchie d'instructions, garde-fous en entrée et en sortie, tests adverses récurrents.
- Injection indirecte et RAG, contrôle et nettoyage des sources avant indexation, application des droits dans la requête vectorielle et non après coup, cloisonnement des index par tenant, signature des sources.
- Fuite et exfiltration, prévention de fuite sur les prompts, classification et masquage des sorties, pseudonymisation sur site avant tout appel externe, interdiction des données sensibles en clair dans le prompt.
- Surexposition en bureautique, revue des droits et étiquettes de sensibilité en amont, recherche restreinte en mesure d'attente, audit de la surexposition avant activation.
- Agence excessive des agents, moindre privilège, identité non humaine dédiée, contrôle d'accès fondé sur l'intention, validation humaine des actions sensibles.
- Consommation non bornée, quotas, limites de débit, budget de jetons, mise en cache et traitement par lots, alertes de coût.

Aucun de ces contrôles n'est universel, on les sélectionne selon les risques retenus en phase 3.

A.5 Fiches détaillées des cas d'usage

Les cas sont anonymisés, aucune organisation n'est identifiable.

Cas A, automatisation low-code avec IA intégrée. Contexte, grand compte, dizaines de scénarios make.com en production, certains appelant un modèle d'IA. Mode, évaluation

de l'existant. Constats, appels d'IA par des clés étrangères au tenant d'entreprise, secrets en clair dans des étapes. Remédiation, bascule de tous les appels vers les API Gemini d'entreprise via Vertex AI avec résidence et journalisation, gestion centralisée des secrets. Reclassification, contrôlé.

Cas B, assistant bureautique en interne. Contexte, assistant d'IA d'entreprise déployé en interne. Constat, certains contenus, notamment classifiés liés à des clients publics ou sensibles, ne doivent pas entrer dans l'outil. Contrôle, restriction par la classe de données, étiquettes de sensibilité, sensibilisation. Reclassification, contrôlé.

Cas C, IA intégrée à un outil de gestion de connaissance. Contexte, Notion déjà gouverné et accepté sur un grand compte, fonction d'IA proposée sur des espaces contenant des données personnelles. Enjeu, le surcroît de risque de la couche IA, hébergement et sous-traitance aux États-Unis, exposition au Cloud Act malgré l'absence d'entraînement et la rétention nulle. Décision, refus de la couche IA en l'état, en attendant une alternative à résidence maîtrisée.

Cas D, assistant souverain en RAG. Contexte, opérateur public de services essentiels refusant toute sortie de données. Mode, construction de zéro. Étude, assistant interne en RAG sur une offre qualifiée SecNumCloud avec modèle à poids ouverts. Points clés, donnée maintenue dans un environnement de confiance, auditabilité complète, maîtrise de l'injection indirecte par contrôle des sources, au prix de l'effort d'ingénierie et d'exploitation.

A.6 Exemple déroulé de bout en bout

Pour montrer la méthode en action, le cas A repris phase par phase, avec ses points de passage. Mode évaluation, l'usage existait déjà.

- Phase 1, cadrage. Des automatisations métier qui résument et routent des documents via un modèle d'IA, sur des données internes confidentielles avec quelques données personnelles. Gate 1 tenu une fois l'usage et la classe de données posés.
- Phase 2, qualification. Plusieurs scénarios appellent l'IA par des clés hors tenant entreprise et des secrets traînent en clair. Gate 2 non tenu, l'accès n'est pas maîtrisé.
- Phase 3, menace. Risques dominants, exfiltration par des clés non maîtrisées et fuite de secrets, à détectabilité faible donc prioritaires. Contrôles exigés, centralisation des secrets et passage forcé par les API d'entreprise.
- Phase 4, architecture. API managée en région UE via le tenant entreprise, avec résidence et journalisation, plutôt qu'un auto-hébergement disproportionné pour ce volume. Gate 4 tenu.
- Phase 5, gouvernance. Reclassification en contrôlé, politique en code, journalisation active, incidents rattachés aux obligations de notification. Gate 5 tenu.

Résultat, une maturité passée de 6 à 13 sur 15 (tableau 9), deux écarts P1 corrigés. En mode intégration, la même grille aurait servi à construire l'automatisation proprement dès le départ.

A.7 Grille de maturité, critères de notation

Chaque phase se note de 0 à 3 selon une échelle commune. 0, absent, rien n'est en place. 1, partiel, traité de façon informelle ou incomplète. 2, maîtrisé, en place et documenté. 3, outillé, en place, documenté et supervisé dans la durée. Le tableau 8 précise, pour chaque phase, ce qui caractérise le niveau maîtrisé, seuil à partir duquel l'usage est considéré sous contrôle. Le niveau 3 ajoute à chaque fois la supervision continue, revue périodique et indicateurs suivis dans le temps.

Phase	Niveau maîtrisé (2) atteint quand
1. Cadrage	L'usage et la classe des données sont décrits et validés par un porteur métier
2. Qualification souveraineté	Résidence, droit applicable et mode d'accès sont documentés et conformes
3. Modélisation de la menace	Les risques majeurs sont identifiés, chacun avec un contrôle ou une acceptation tracée
4. Architecture	La configuration est choisie, argumentée et cohérente avec la classe de données
5. Gouvernance et supervision	Politique écrite, journalisation active et rattachement aux obligations d'incident

TABLE 8. Critères du niveau maîtrisé par phase.

A.8 Exemple de grille de maturité

Application de la grille de maturité (notation de 0 à 3 par phase) au cas A avant remédiation.

```
Phase 1 cadrage de l'usage          : 2
Phase 2 qualification souverainete  : 1
Phase 3 modelisation de la menace   : 0
Phase 4 architecture                : 2
Phase 5 gouvernance et supervision  : 1
Score global = 6 / 15
```

Lecture, un usage en production mais sans modélisation de menace ni supervision, profil à risque courant. La remédiation a visé en priorité les phases 3 et 5.

A.9 Méthode de notation et détail des scores de la campagne

Les scores du corps ne sont pas posés à la main, ils découlent de la grille. Chaque usage a été noté phase par phase selon les critères du tableau 8, à partir de ce qui était réellement observé en mission, la configuration des outils, la revue des scénarios et des permissions, et les échanges avec les équipes. Chaque phase vaut de 0 à 3, le total sur 15. Le tableau 9 donne le détail avant et après remédiation pour chaque usage, ce qui rend chaque chiffre du corps traçable jusqu'à son origine.

Usage	Avant (P1..P5)	Total	Après (P1..P5)	Total
Assistant bureautique (interne)	1, 1, 0, 1, 1	4	3, 2, 2, 2, 2	11
Gemini d'entreprise	2, 1, 0, 1, 1	5	3, 2, 2, 2, 2	11
make.com + Vertex	2, 1, 0, 2, 1	6	3, 3, 2, 2, 3	13
AppSheet avec IA	1, 1, 0, 1, 0	3	2, 2, 2, 2, 1	9
Notion IA	2, 1, 0, 1, 0	4	refusé	-

TABLE 9. Détail des notations par phase, avant et après remédiation. P1 cadrage, P2 qualification souveraineté, P3 menace, P4 architecture, P5 gouvernance.

Une régularité ressort, la phase 3, la modélisation de la menace, est presque toujours à zéro avant intervention, et la phase 5, la gouvernance et la supervision, à peine amorcée. C'est le profil type d'un usage activé par le métier sans passage par la sécurité, et c'est exactement ce que la remédiation corrige en priorité.

A.10 Arbre de décision

L'arbre de la figure 6 résume la logique du cadre, depuis la question qui peut tout arrêter, la sensibilité de la donnée, jusqu'à la mise en production.

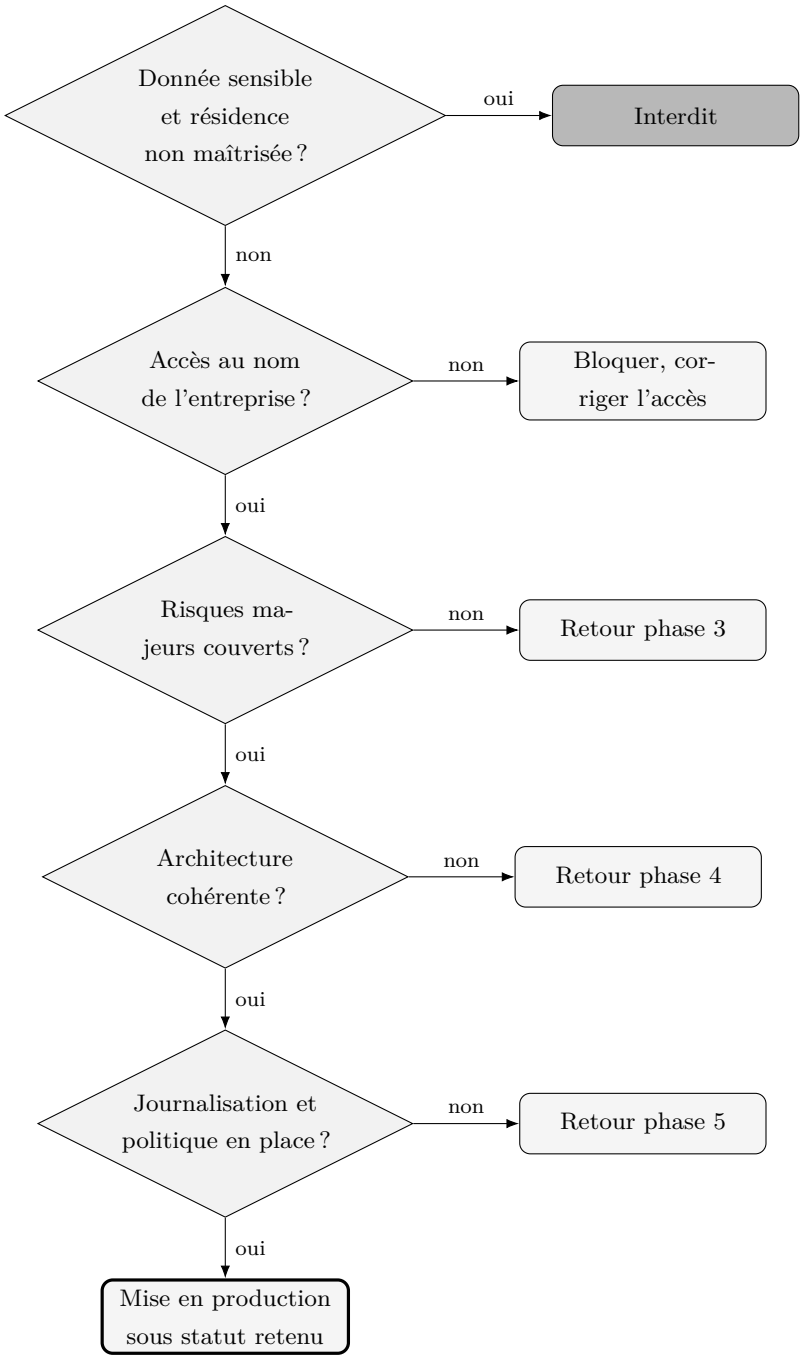

FIGURE 6. Arbre de décision synthétique du cadre.

A.11 Modèle de coût des architectures

Ordres de grandeur relevés en juin 2026, pour donner corps à l'arbitrage du focus 2. Les tarifs d'API sont en dollars (fournisseurs américains) et le coût GPU en euros (fournisseur européen), du même ordre au change actuel. Les prix évoluent vite, ils valent comme structure de raisonnement, pas comme devis.

Option	Coût indicatif	Remarques
API managée, modèle haut de gamme	env. 5 $ entrée et 25 $ sortie par million de jetons [2]	Traitement par lot −50%, cache de prompt jusqu'à −90%
API managée, modèle européen	env. 2 $ entrée et 6 $ sortie par million de jetons [11]	Réversibilité plus simple vers l'auto-hébergement
Souverain auto-hébergé	env. 2,73 € par heure et par GPU H100 [16]	Coût fixe, pas de coût au jeton, exige de l'ingénierie

TABLE 10. Coûts indicatifs des configurations, juin 2026.

Le point de bascule est moins favorable à l'auto-hébergement qu'un calcul naïf ne le laisse croire, et il faut le poser honnêtement. Un modèle de soixante-dix milliards de paramètres servi sur un nœud de quatre H100 avec un moteur d'inférence optimisé atteint de l'ordre de trois mille jetons par seconde. À pleine charge, cela représenterait près de huit milliards de jetons par mois, mais aucun service ne tourne à pleine charge, et à un taux d'utilisation réaliste d'environ trente pour cent, on retombe autour de deux à trois milliards de jetons par mois. Ce nœud coûte près de huit mille euros par mois, auxquels s'ajoutent une redondance pour ne pas dépendre d'une seule machine, ce qui double la facture, puis le coût d'ingénierie et de MLOps, supervision, mises à jour, sécurisation, soit une fraction de temps plein loin d'être négligeable.

Mis bout à bout, à ces volumes réalistes de deux à trois milliards de jetons par mois, l'API managée reste nettement moins chère que le nœud souverain une fois celui-ci redondé, et avant même de compter l'ingénierie. L'auto-hébergement ne s'en rapproche qu'à très forte charge soutenue, quand un nœud tourne en continu près de sa capacité, ce qui reste rare. La conséquence est nette, et c'est l'inverse d'une intuition répandue, on retient le souverain pour la souveraineté, l'auditabilité ou la réversibilité, presque jamais pour le seul prix.

A.12 Réversibilité, schéma de la stratégie de sortie

La figure 7 schématise la voie de sortie d'un usage critique, depuis une API managée vers une cible de repli souveraine, avec ses déclencheurs et son délai indicatif. C'est la traduction technique de la stratégie de sortie attendue par DORA.

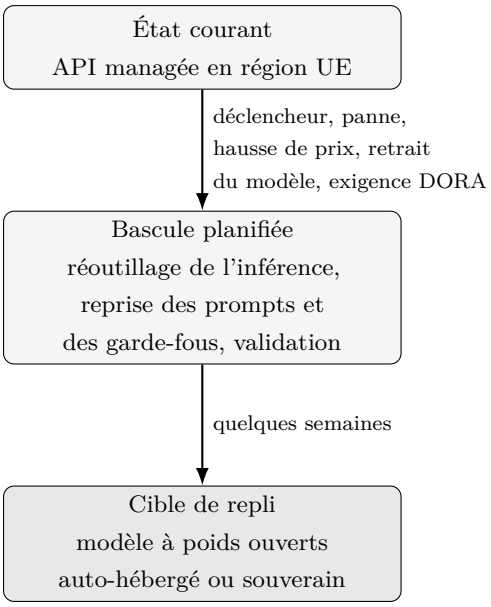

FIGURE 7. Voie de sortie d'un usage critique, de l'API managée vers le souverain.

A.13 Correspondance du cadre avec les référentiels

Le cadre n'invente pas une théorie, il opérationnalise des référentiels existants. Le tableau 11 trace cette correspondance, phase par phase.

Phase du cadre	NIST AI RMF	ISO/IEC 42001	ANSSI IA gén.	OWASP LLM	AI Act
1. Cadrage et classification	Map	Contexte, objectifs	Cadrage de l'usage	-	Niveau de risque
2. Qualification souveraineté	Govern, Map	Fournisseurs, données	Hébergement, données	LLM03, LLM02	Résidence, transparence
3. Modélisation menace	Measure	Appréciation du risque	Scénarios d'attaque	Top 10 complet	Gestion du risque
4. Architecture	Manage	Ressources	Déploiement	LLM03, LLM10	Robustesse
5. Gouvernance, supervision	Govern, Manage	Surveillance, amélioration	Production, journaux	LLM05, LLM06	Surveillance post-marché

TABLE 11. Correspondance des cinq phases avec les principaux référentiels.

A.14 Calendrier de l'AI Act

La figure 8 récapitule l'application échelonnée de l'AI Act, telle que modifiée par le Digital Omnibus de mai 2026. Elle aide à situer un usage dans le temps réglementaire plutôt qu'à le rapporter à une échéance unique. Dans le détail, l'Omnibus reporte les obligations des systèmes à haut risque de l'annexe III du 2 août 2026 au 2 décembre 2027, celles de l'annexe I au 2 août 2028, et le marquage des contenus de l'article 50(2) au 2 décembre 2026, tout en ajoutant à l'article 5 une interdiction des nudifiers et du matériel pédocriminel. Les autres obligations de transparence, y compris côté déployeur, et la littératie restent applicables au 2 août 2026. L'accord reste provisoire, sous réserve de son adoption définitive et de sa publication au Journal officiel.

FIGURE 8. Calendrier d'application de l'AI Act, intégrant les reports du Digital Omnibus de mai 2026.

A.15 Niveaux de souveraineté

La souveraineté n'est pas binaire. Le tableau 12 en distingue trois niveaux, avec leurs contrôles compensatoires possibles.

Niveau	Résidence	Droit applicable	Contrôles compensatoires
Souverain qualifié (type SecNumCloud)	UE	UE	Peu nécessaires
Hyperscaler en région UE	UE	Exposé au Cloud Act	Pseudonymisation sur site, chiffrement client, clés gérées par le client
SaaS ou cloud global	Hors UE possible	Étranger	Souvent insuffisants pour données sensibles

TABLE 12. Trois niveaux de souveraineté et leurs contrôles compensatoires.

A.16 Schéma de flux et frontières de confiance

Le schéma de la figure 9 modélise un usage type, un assistant interrogeant des sources internes via une passerelle de contrôle avant d'appeler un modèle externe. Les frontières de confiance, en pointillés, sont les endroits où la donnée change de domaine de maîtrise, et donc où se concentrent les contrôles.

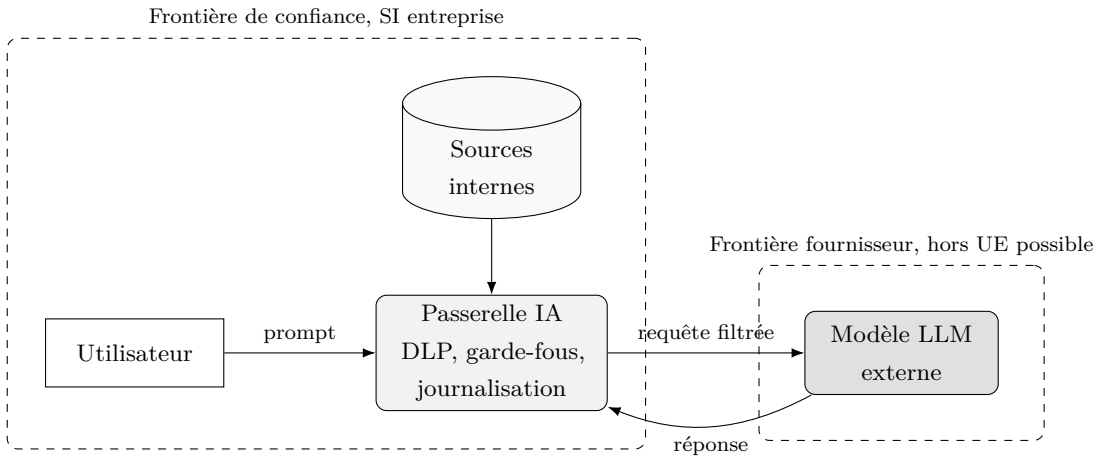

FIGURE 9. Flux de données et frontières de confiance d'un usage type.

A.17 Politique-as-code, exemple complet

Version étendue de l'ossature présentée dans le corps du mémoire, pour un usage contrôlé.

```
usage: assistant-support-client
classe_donnees: dcp_limitees
statut: controle
famille: T3
fournisseur:
  nom: modele-ue-manage
  residence: eu
  droit_applicable: ue
  entrainement_sur_donnees: false
  retention_jours: 0
controles:
  entree:
    dlp: true
    pseudonymisation_sur_site: true
    prompts_autorises: liste_blanche_support
  sortie:
    classification: true
    masquage_dcp: true
    verification_humaine: si_envoi_externe
  journalisation:
    actif: true
    contenu: [horodatage, utilisateur, hash_prompt, categorie_sortie, decision_garde_fou]
    retention_jours: 365
supervision:
  revue: trimestrielle
  kpi: [taux_blocage_dlp, incidents, volume, taux_hallucination_signale]
incident:
  schemas:
    - fuite_dcp_par_sortie
    - injection_indirecte_avere
    - indisponibilite_fournisseur
```

```
rattachement_nis2:
  alerte_precoce_h: 24
  notification_h: 72
  rapport_final_jours: 30
```

A.18 Structure de l'étude complémentaire

Le questionnaire en ligne, toujours ouvert, est organisé en sept blocs, pour suivre le cheminement réel d'une organisation face à l'IA.

1. Le moment déclencheur, quand l'IA est devenue un sujet cyber et avec quel cadre.
2. Arbitrages et tensions, performance contre sécurité et souveraineté, shadow AI, incidents.
3. Architecture et choix techniques, modèles ouverts, maillon le plus à risque, visibilité sur les API.
4. Gouvernance et régulation, qui décide, état sur l'AI Act, frameworks de référence, budget.
5. L'IA au service de la cybersécurité, outils, paradoxe du LLM étranger pour protéger un SI européen.
6. Technologies émergentes et trajectoire, agents, modèles locaux, RAG, IA embarquée.
7. Vision et clôture, scénario à trois ans, risque sous-évalué, conseils.

Les résultats consolidés seront publiés sur research.thomas.how.

A.19 Points de comparaison auprès de RSSI, par secteur

Synthèse anonymisée des échanges avec des RSSI français et suisses de plusieurs secteurs, sur ce qu'ils avaient en place et leur façon de traiter le risque IA. Aucune organisation n'est identifiable.

Secteur	Posture de départ	Point fort observé	Sensibilité dominante
Finance (FR et CH)	Encadrement déjà amorcé	Risque fournisseur et sortie, tirés par DORA et la FINMA	Confidentialité, traçabilité
Énergie	Restrictive et prudente	Cloisonnement, exigence de souveraineté	Résidence, continuité, NIS2
Industrie	Adoption par les usages	Rapidité de déploiement	Shadow AI, propriété intellectuelle

Table 13. Pratiques de gestion du risque IA observées chez des RSSI rencontrés, par secteur.

Transverse aux trois secteurs, une spécificité suisse revient nettement, la résidence des données sur le territoire et la nLPD posées d'emblée, avec une culture forte de confidentialité.

Deux constats traversent ces échanges. La donnée et sa résidence reviennent partout comme le vrai point de décision, ce qui conforte le choix d'en faire la première phase du cadre. Et la

dépendance au fournisseur est jugée sous-estimée dans tous les secteurs, ce qui justifie d'avoir hissé la réversibilité au rang d'exigence d'architecture.