AIDES
A LA REFLEXION

Déléguer, assister, sanctuariser

Quel modèle de sécurité apprenez-vous à l'IA ?

Alain Quiot

Rédaction coordonnée par Eric Marsden

n° 2026-03

THÉMATIQUE

Transition numérique

Les Cahiers de la sécurité industrielle

FONCSI
Fondation pour une culture de sécurité industrielle

La *Fondation pour une Culture de Sécurité Industrielle* (Foncsi) est une Fondation de recherche reconnue d'utilité publique par décret en date du 18 avril 2005. Elle a pour ambitions de :

- ▷ contribuer à l'amélioration de la sécurité dans les entreprises industrielles de toutes tailles, de tous secteurs d'activité ;
- ▷ rechercher, pour une meilleure compréhension mutuelle et en vue de l'élaboration d'un compromis durable entre les entreprises à risques et la société civile, les conditions et la pratique d'un débat ouvert prenant en compte les différentes dimensions du risque ;
- ▷ favoriser l'acculturation de l'ensemble des acteurs de la société aux problèmes des risques et de la sécurité.

Pour atteindre ces objectifs, la Fondation favorise le rapprochement entre les chercheurs de toutes disciplines et les différents partenaires autour de la question de la sécurité industrielle : entreprises, collectivités, organisations syndicales, associations. Elle incite également à dépasser les clivages disciplinaires habituels et à favoriser, pour l'ensemble des questions, les croisements entre les sciences de l'ingénieur et les sciences humaines et sociales.

Fondation pour une Culture de Sécurité Industrielle
Fondation de recherche, reconnue d'utilité publique

www.FonCSI.org

6 allée Émile Monso – CS 22760
31077 Toulouse cedex 4
France

Courriel : contact@FonCSI.org

Abstract

Title	Delegate, assist, ring-fence: which safety model are you teaching your AI?
Keywords	AI, safety model, automation, delegation, subcontracting, experience feedback
Authors	Alain Quiot
Publication date	June 2026

AI is entering high-risk industries with a twofold promise: handling the informational complexity that traditional tools can no longer master, and detecting signals in the data that no one sees. This document examines that promise from viewpoint of a practitioner, rather than that of a researcher or technologist. Its starting point is a diagnosis: six structural weaknesses affect high-risk industries: complexity of the regulatory framework, preparation disconnected from risk control, fragmentation of the extended enterprise, weak signals left unassembled, operational experience feedback that fails to question the safety model, and indicators centred on activity rather than on the robustness of the system. Each has an informational dimension that AI can help to address. But before deploying, one question — which no one seems to have asked — must be answered: which safety model are you teaching the machine? The document proposes a decision grid — delegate, assist, ring-fence — broken down into six operational questions, and shows that the choice of the level of assistance (the human drives the AI, the AI advises the human, or the AI produces and the human supervises) shapes the impacts on skills, vigilance, shared understanding, accountability and safety culture. It identifies a specific organizational risk — accountability laundering, whereby the human signature lends a veneer of governance to what is in fact an algorithmic decision — and proposes several countermeasures, beginning with the construction of a protected space for contestation.

The document closes with questions for an executive committee and with a debate on unresolved tensions — between transparency and performance, between assistance and deskilling, between data sharing and data protection.

About the authors

Alain Quiot has more than forty years of experience in industrial safety, primarily in the nuclear sector at EDF. Through his involvement in the program to redesign the information system of the nuclear fleet, he became aware of the gap between the wealth of data available within an industrial organization and the actual ability to cross-reference, interpret, and transform that data into safety-related decisions. This experience sparked his interest in AI, not as just another technological promise, but as a tool potentially capable of addressing the informational dimension of the vulnerabilities he had observed throughout his career. From 2023 to 2026, he joined the Icsi as an advisor to the director.

To cite this document

Quiot, A. (2026), *Delegate, assist, ring-fence: which safety model are you teaching your AI?* Number 2026-03 of the *Cahiers de la Sécurité Industrielle*, Foundation for an Industrial Safety Culture, Toulouse, France (ISSN 2100-3874). DOI: 10.57071/iaq420. Available from FonCSI.org/en.

Résumé

Titre	Déléguer, assister, sanctuariser : quel modèle de sécurité apprenez-vous à l'IA ?
Mots-clefs	IA, modèle de sécurité, automatisation, délégation, entreprise étendue, REX
Auteurs	Alain Quiot
Date de publication	juin 2026

L'IA arrive dans les industries à haut risque avec une double promesse : traiter la complexité informationnelle que les outils traditionnels ne maîtrisent plus, et détecter dans les données des signaux que personne ne voit.

Ce *Cahier* examine cette promesse du point de vue du praticien de la sécurité industrielle — pas du chercheur ni du technologue. Son point de départ est un diagnostic : six fragilités structurelles traversent les industries à haut risque — complexité du référentiel, préparation découplée de la maîtrise des risques, fragmentation de l'entreprise étendue, signaux faibles non assemblés, retour d'expérience qui ne questionne pas le modèle de sécurité, indicateurs centrés sur l'activité plus que sur la robustesse du système. Chacune comporte une dimension informationnelle que l'IA peut aider à traiter. Mais avant de déployer, il faut répondre à une question que personne ne semble avoir posée : quel modèle de sécurité enseignez-vous à la machine ? Le *Cahier* propose une grille de décision — déléguer, assister, sanctuariser — déclinée en six questions opérationnelles, et montre que le choix du niveau d'assistance (l'humain pilote l'IA, l'IA conseille l'humain, ou l'IA produit et l'humain supervise) détermine les impacts sur les compétences, la vigilance, la compréhension partagée, la responsabilité et la culture de sécurité. Il identifie un risque organisationnel spécifique — le blanchiment de légitimité, où la signature humaine donne une apparence de gouvernance à une décision effectivement algorithmique — et propose plusieurs parades, à commencer par la construction d'un espace de contestation protégé.

Le Cahier se conclut par des questions pour un Codir et par une mise en débat des tensions — entre transparence et performance, entre assistance et affaiblissement, entre partage et protection des données.

À propos des auteurs

Alain Quiot possède plus de quarante ans d'expérience en sécurité industrielle, principalement dans le nucléaire chez EDF. En participant au programme de refonte du système d'information du parc nucléaire, il a mesuré la distance entre la richesse des données disponibles dans une organisation industrielle et la capacité réelle à les croiser, les interpréter et les transformer en décisions de sécurité. C'est cette expérience qui a fondé sa curiosité pour l'IA — non pas comme une promesse technologique de plus, mais comme un outil potentiellement capable de traiter la dimension informationnelle des fragilités qu'il avait observées pendant toute sa carrière. De 2023 à 2026, il a intégré l'Icsi en tant que chargé de mission auprès du directeur général.

Courriel : `alain.quiot@edf.fr`

Pour citer ce document

Quiot, A. (2026), *Déléguer, assister, sanctuariser : quel modèle de sécurité apprenez-vous à l'IA ?* Numéro 2026-03 des *Cahiers de la Sécurité Industrielle*, Fondation pour une Culture de Sécurité Industrielle, Toulouse, France (ISSN 2100-3874). DOI : 10.57071/aiq420. Disponible à l'adresse `FonCSI.org/fr`.

Table des matières

Introduction

Contexte

L'IA entre dans les industries à haut risque. Elle y entre par la maintenance prédictive, par l'analyse automatisée des rapports d'incidents, par la vision par ordinateur sur les chantiers, par les assistants conversationnels qui aident les opérateurs à naviguer dans des milliers de pages de documentation. Elle y entre parce que les éditeurs la proposent, parce que les directions la demandent, et parce que les problèmes informationnels des organisations industrielles — données dispersées, signaux non croisés, complexité croissante — appellent des outils que les approches traditionnelles n'offrent pas.

Mais l'IA n'arrive pas seulement comme une fonctionnalité ajoutée aux outils existants. Elle arrive aussi par une reconfiguration plus profonde des architectures de données industrielles : une partie de la capacité à détecter, corréler, diagnostiquer et recommander se déplace des outils métier traditionnels — GMAO, plateformes QHSE, gestion documentaire — vers des couches transverses qui relient les flux issus des installations, les historiques de maintenance, les documents techniques, les alarmes et les modèles d'actifs. Il ne s'agit donc plus seulement d'informatiser un processus métier, mais d'organiser une capacité d'apprentissage à partir de données industrielles hétérogènes.

Ce déplacement est décisif pour les industries à haut risque. Il peut rendre visibles des signaux jusque-là dispersés : une dérive de procédé, un historique de maintenance dégradé, un retour d'expérience oublié, une alarme chronique, une fragilité documentaire. Mais il crée aussi un risque nouveau : produire une intelligence industrielle puissante sans modèle de sécurité explicite. Une plateforme pourrait apprendre à optimiser la disponibilité, réduire les temps d'arrêt ou fluidifier la maintenance sans savoir distinguer ce qui relève d'un arbitrage acceptable, d'une érosion de marge ou d'une fragilisation de barrière critique. La question n'est donc plus seulement : que sait faire l'IA ? Elle devient : quelle architecture d'apprentissage sommes-nous en train de construire, et autour de quel modèle de sécurité ?

D'après l'auteur, les industries à risque n'ont pas un problème d'IA — elles ont des problèmes que l'IA pourrait aider à traiter, à condition de savoir ce que l'on confie à la machine, ce que l'on garde sous contrôle humain, et ce que l'on protège de toute interférence algorithmique. Cette question doit être posée avant le déploiement, pour maîtriser les effets sur les compétences, la vigilance et la culture de sécurité.

Objectifs du document

L'objectif de ce document est de donner au praticien de la sécurité industrielle — directeur de site, directeur HSE, responsable d'exploitation — des clés pour poser ces bonnes questions avant de décider. Il ne s'agit ni de promouvoir l'IA ni de la freiner. Il s'agit de penser son intégration avec la même rigueur que celle que les industries à risque appliquent à toute modification de leurs installations : en explicitant ce que l'on veut protéger, en évaluant les effets sur l'humain et sur l'organisation, en prévoyant les modes dégradés, et en préservant la capacité de contester ce que la machine propose.

Ce *Cahier* est écrit par un praticien, pour des praticiens. Il ne prétend ni à l'exhaustivité technique, ni à la neutralité académique : l'auteur a des convictions, nourries par l'expérience du terrain, et il les assume. La posture revendiquée est celle de **l'anticipation éclairée** : questionner les promesses de l'IA sans les rejeter, nommer les risques sans les dramatiser, et proposer des repères pour décider — pas des réponses définitives.

Structure du document

Dans ce document, le terme « IA » ne recouvre pas un objet unique : caméras intelligentes, algorithmes prédictifs, assistants conversationnels, plateformes de données augmentées ne posent pas les mêmes questions de sécurité. Le préambule précise cette typologie.

Avant de se demander ce que l'IA peut faire, il faut regarder en face ce que les organisations à risque n'arrivent pas à résoudre complètement, malgré des décennies d'efforts et d'investissements considérables. Six fragilités structurelles traversent ces organisations : complexité du référentiel, préparation parfois découplée du retour d'expérience, fragmentation de l'entreprise étendue, signaux faibles non assemblés, apprentissage organisationnel perfectible, indicateurs mal alignés avec la robustesse des défenses. Chacune comporte une dimension informationnelle que l'IA peut aider à traiter — mais seulement à condition d'avoir préalablement explicité le modèle de sécurité que l'on veut faire apprendre à la machine. C'est l'objet du chapitre 1.

Le chapitre 2 décrit les différentes familles d'IA (apprentissage machine, réseaux de neurones profonds, vision par ordinateur, grands modèles de langage, jumeaux numériques, IA symbolique, réalité virtuelle), ce que ces systèmes savent faire et ce qu'ils ne savent pas faire. Le chapitre 3, le cœur du document, traite des données sur lesquelles travaillera l'IA, et le modèle de sécurité associé. Il propose une classification des types de décisions qui devraient rester sous contrôle humain, protégées d'un traitement algorithmique, et celles qui peuvent être déléguées à la machine, avec la grille « déléguer - assister - sanctuariser » proposée par l'auteur.

Le chapitre 4 propose six questions opérationnelles pour montrer concrètement ce que l'IA peut apporter, et à quelles conditions. Les questions sont formulées du point de vue du dirigeant et du directeur HSE. Le chapitre 5 applique l'analyse aux relations organisationnelles dans l'entreprise étendue, qui pose des problèmes spécifiques de propriété de données et de silence organisationnel.

Le chapitre 6 propose un ensemble de questions qu'un comité de direction devrait avoir traitées avant de considérer que son organisation est prête à déployer l'IA. L'auteur plaide pour une introduction progressive de l'IA, par paliers. Le dernier chapitre ouvre un débat sur cinq tensions que produit l'introduction de système d'IA, liées à des choix de valeur, avec des éclairages pour faciliter un arbitrage adapté à chaque contexte organisationnel.

En annexe, le lecteur intéressé trouvera un glossaire des termes techniques et acronymes utilisés dans le document, une illustration d'application de la grille déléguer-assister-sanctuariser, une cartographie des familles d'IA et des sources de données adaptées à chaque famille, et enfin un résumé d'un document *Considerations for Developing Artificial Intelligence Systems in Nuclear Applications* proposé par trois organismes de contrôle du secteur nucléaire.

Comment l'IA a été utilisée pour ce Cahier

Le sujet même de ce document rend nécessaire d'expliciter comment l'IA a contribué à sa production. Des outils ont été mobilisés : Perplexity (Perplexity AI) et Semantic Scholar pour explorer la littérature scientifique ; NotebookLM (Google) pour cartographier et dialoguer avec les publications retenues ; Claude AI (Anthropic) pour structurer l'argumentation, reformuler et tester la cohérence interne du texte.

En appliquant la grille DAS qui sert d'épine dorsale au document, ces outils relèvent tous du niveau d'*assistance outil* : aucun n'a été placé en position de *délégation*, c'est-à-dire de produire un signal accepté tel quel. Des zones ont été explicitement *sanctuarisées* : le choix des thèses défendues, et l'arbitrage final sur ce qui mérite d'être écrit. Les références bibliographiques ont été vérifiées en source primaire. La responsabilité du texte incombe pleinement à l'auteur.

Préambule : de quelle IA parlons-nous ?

Le terme d'« IA » dans ce document ne se réfère pas à un objet unique, mais plutôt à une mosaïque de technologies, d'usages et d'architectures. Les sites industriels accueillent aujourd'hui des systèmes à base d'IA qui n'ont rien en commun, ni dans leur conception, ni dans leur usage, ni dans leurs effets sur la sécurité industrielle : caméras intelligentes, algorithmes de maintenance prédictive, assistants conversationnels, jumeaux numériques, plateformes de données industrielles capables de croiser des flux issus de la supervision, de la maintenance, de la documentation ou du terrain. Pour éviter une confusion fréquente, il faut distinguer deux questions :

- La première est technologique : *de quel type d'IA parle-t-on ?* Vision par ordinateur, apprentissage automatique, modèles prédictifs, grands modèles de langage, systèmes de recommandation, optimisation, agents logiciels, IA générative ?
- La seconde est organisationnelle : *où cette IA s'insère-t-elle dans le système industriel, et quel effet peut-elle avoir sur la manière dont l'organisation voit, interprète, décide et apprend ?*

Ce document privilégie la seconde question. Non parce que la première serait secondaire, mais parce que le risque principal, en sécurité industrielle, ne tient pas seulement à la performance d'un modèle algorithmique. Il tient à l'usage qui en est fait, au processus dans lequel il s'insère, aux décisions qu'il influence, aux compétences qu'il déplace.

Sans prétendre à une typologie exhaustive, trois registres d'intégration de l'IA doivent être distingués :

- **Les IA spécialisées**, embarquées dans des outils, des équipements ou des dispositifs techniques : caméras intelligentes pour la vidéosurveillance, algorithmes de reconnaissance de défauts intégrés aux contrôles non destructifs, capteurs autonomes de détection, drones ou robots d'inspection assistés par IA. Ces systèmes sont aujourd'hui largement déployés. Ils sont conçus et entraînés par des fournisseurs, souvent vendus avec l'équipement ou le service, et utilisés dans un niveau d'assistance d'outil au sens classique : ils étendent une capacité perceptive, analytique ou instrumentale humaine, sans reconfigurer en profondeur l'organisation du travail. Ces IA portent un modèle local — de reconnaissance, de classification, de mesure ou de détection — et non un modèle global de sécurité de l'organisation. Une caméra détecte une présence en zone interdite ; un algorithme de contrôle non destructif classe un signal comme suspect ; un capteur prédictif signale une dérive. Ces dispositifs peuvent contribuer à la sécurité, mais ils ne définissent pas, à eux seuls, ce qu'une organisation considère comme sûr, acceptable, prioritaire ou contestable. Ils ne constituent pas l'objet principal de ce document.
- Les **IA applicatives intégrées aux systèmes métier** existants : GMAO, suites d'asset management, plateformes QHSE, outils documentaires, bases de REX, systèmes de planification, applications de conduite ou de supervision. Dans ce registre, l'IA vient enrichir des outils que les organisations utilisent déjà pour préparer les interventions, suivre les écarts, gérer les plans d'action, documenter les contrôles ou piloter la conformité. Elle peut aider à rédiger, synthétiser, rechercher, classer, recommander, détecter des écarts, rapprocher des cas similaires, qualifier un événement ou proposer un ordre de travail. Son apport potentiel est important : rendre plus accessibles des informations dispersées, réduire le temps de recherche documentaire, faciliter la mobilisation du retour d'expérience, améliorer la qualité des historiques d'intervention. Sa limite l'est tout autant : ces IA restent structurées par les objets que les systèmes métier savent déjà gérer — équipements, ordres de travail, écarts, audits, documents, versions. Elles ne suffisent pas, à elles seules, à rendre visibles les boucles qui relient ces objets entre eux.
- Les **plateformes de données industrielles augmentées par IA** : ce registre, le plus récent et le plus structurant, rassemble des plateformes de données industrielles qui agrègent des flux issus de sources hétérogènes — installations, historiques de maintenance,

documents techniques, alarmes, images, modèles d'actifs, données d'entreprise — et qui s'augmentent progressivement de capacités d'IA pour détecter, corréler, diagnostiquer et recommander. Ces plateformes ne sont pas une simple extension des systèmes métier. Elles déplacent une partie de la capacité d'analyse vers une couche transverse, capable de croiser ce que les outils traditionnels ne pouvaient pas relier : une dérive de procédé avec un historique de maintenance dégradé, un retour d'expérience oublié avec une configuration d'intervention à risque, une alarme chronique avec une fragilité documentaire.

C'est ce registre qui pose les questions les plus nouvelles pour la sécurité industrielle. Parce qu'il ne s'agit plus d'informatiser un processus métier, mais d'organiser une capacité d'apprentissage à partir de données hétérogènes — sans que le modèle de sécurité de l'organisation soit nécessairement inscrit dans cette capacité d'apprentissage. C'est principalement ce registre qui est traité dans le présent document.

La place particulière des grands modèles de langage (LLM)

Une précision est nécessaire concernant les grands modèles de langage — LLM — parce qu'ils constituent aujourd'hui la forme d'IA la plus visible pour beaucoup de décideurs et de praticiens. Pour nombre d'utilisateurs, l'IA désigne d'abord ChatGPT, Gemini, Claude, Mistral ou leurs équivalents intégrés dans les outils professionnels. Cette perception est compréhensible, mais elle peut devenir trompeuse si elle conduit à réduire l'IA industrielle aux seuls assistants conversationnels.

Les LLM ne constituent pas un quatrième registre. Ils sont une technologie transverse qui peut être intégrée dans chacun des trois registres précédents : dans un outil métier pour interroger une base documentaire ou synthétiser un retour d'expérience ; dans une plateforme de données pour interroger en langage naturel des historiques de maintenance ou des alarmes ; ou en interface d'un système plus spécialisé, en traduisant une question humaine en requête technique. Leur force et leur fragilité tiennent au même endroit.

- **Force : leur capacité à travailler le langage** — lire, résumer, reformuler, classer, rapprocher, structurer une argumentation.
- **Fragilité : ils peuvent produire une réponse plausible sans que celle-ci soit vraie**, complète, contextualisée ou applicable.

Dans les industries à haut risque, un LLM ne devrait donc jamais être défini seulement comme un « assistant conversationnel ». Il devrait être qualifié par son régime d'usage : aide à la recherche, aide à l'analyse, aide à la rédaction, aide au diagnostic, aide à la décision, ou production d'un livrable soumis à validation humaine.

Un risque : le shadow AI

Une dernière raison de se méfier d'une définition trop étroite des LLM tient à leur mode actuel de pénétration dans les organisations. Les enquêtes récentes — MIT *State of AI in Business 2025*, Microsoft *Work Trend Index*, Netskope *Cloud and Threat Report 2025* — convergent sur un même constat : entre 60 et 90% des collaborateurs utilisent des LLM dans leur travail, mais une fraction beaucoup plus faible le fait avec une autorisation explicite et un cadre formalisé.

Les organisations à haut risque ne sont pas exemptes de ce phénomène, désormais désigné sous le nom de *shadow AI*. Pour le praticien de la sécurité industrielle, ce point est central : avant même qu'une organisation ne décide *si* et *comment* déployer l'IA, celle-ci est déjà à l'œuvre, par la porte de derrière, sur des activités qui peuvent toucher à la préparation, à l'analyse de retour d'expérience ou à la rédaction de documents techniques. On y revient au chapitre 3.

1

Six fragilités que l'IA oblige à regarder en face

Point clé

Les difficultés les plus déterminantes pour l'exploitation d'une installation industrielle à haut risque ne sont pas des défaillances isolées. Ce sont des fragilités structurelles du système sociotechnique — de sa conception, de son pilotage, de ses interfaces. Ces fragilités affectent simultanément la sûreté de fonctionnement, la prévention des accidents graves et mortels, la disponibilité des installations et, par conséquent, la performance économique.

Ce chapitre identifie six fragilités systémiques qui traversent les industries à haut risque et qui, malgré trente ans de progrès en gestion de la sécurité, continuent de résister aux solutions classiques. Les six principales fragilités proposées ici sont nées dans le secteur nucléaire. C'est le terrain que l'auteur connaît le mieux, celui où il a exercé pendant plus de trente-cinq ans. Trois années passées à l'Icsi, au contact d'exploitants de la chimie, de l'*oil and gas*, du ferroviaire… l'ont convaincu que ces fragilités ne sont pas propres à une industrie spécifique. Elles résonnent dans beaucoup de secteurs industriels, avec des intensités variables et des modalités différentes.

L'auteur ne prétend pas à la neutralité sectorielle. Ce chapitre est un outil de diagnostic, que le lecteur est invité à confronter à sa propre réalité. Si trois des six fragilités lui parlent davantage que les trois autres, c'est déjà un résultat utile.

Une précision est nécessaire avant d'entrer dans le diagnostic. Les six fragilités décrites dans ce chapitre ne sont pas des découvertes. Elles sont documentées depuis des années dans des rapports publics — de l'ASN, de l'AIEA, de l'Ineris, de l'IOGP… et l'auteur a utilisé ces sources publiques pour les construire.

Le *Cahier* de la Foncsi sur la sécurité à l'ère du « vivre avec » **[Bieder et al. 2024]** montre que ces fragilités structurelles s'inscrivent dans un contexte plus large de complexification et d'incertitude croissante — changement climatique, digitalisation, tensions géopolitiques, évolution du rapport au travail — qui met à l'épreuve les paradigmes fondateurs du management de la sécurité. Les fragilités décrites ici sont à la fois des problèmes d'aujourd'hui et des amplificateurs de ce qui vient. Les industriels les connaissent. Ils les travaillent, parfois depuis plus d'une décennie, avec des résultats inégaux, mais réels. Les programmes de simplification, les dispositifs de préparation renforcée, les architectures de REX, les réformes d'indicateurs — tout cela existe et progresse.

Ce qui est proposé ici n'est donc pas un réquisitoire contre les industriels ni un constat d'échec. C'est un regroupement structuré de difficultés connues, mis en forme dans une perspective précise : comprendre pourquoi ces problèmes résistent malgré les efforts déployés, et identifier ce que cette résistance a de spécifiquement informationnel — ce qui ouvrira, dans les chapitres suivants, la question de l'apport possible de l'IA.

Le fait que ces fragilités persistent ne signifie pas que les organisations sont défaillantes. Cela signifie qu'elles affrontent des problèmes structurels dont la résolution dépasse la seule volonté d'un industriel — parce qu'ils impliquent des régulateurs, des législateurs, des chaînes de fournisseurs, des temporalités longues et une complexité intrinsèque aux systèmes sociotechniques à haut risque.

1.1 La complexité du référentiel dépasse la capacité d'appropriation

FRAGILITÉ 1

Les industries à haut risque ne manquent pas de règles. Elles en ont beaucoup, et ces règles s'accumulent sans que les anciennes soient suffisamment simplifiées, hiérarchisées ou retirées. Dans le secteur nucléaire, les règles générales d'exploitation, les prescriptions issues des réexamens décennaux, les référentiels de maintenance, les dossiers d'essais et les retours d'expérience forment une architecture documentaire dont la densité et le volume sont devenus un problème en soi.

Dans les ICPE Seveso, le système de gestion de la sécurité, les études de dangers et les plans d'urgence imposent un référentiel dense aux exploitants de seuil haut. Le Barpi identifie régulièrement, dans son analyse des causes profondes des accidents industriels, que l'accès à l'information pertinente est entravé par le cloisonnement entre services — production, maintenance, sécurité, environnement — et identifie le décloisonnement comme un critère central de qualité organisationnelle[1]. Dans le pétrole et le gaz, l'IOGP gère plus de deux cents rapports techniques pour la seule sécurité, nécessitant des efforts constants d'harmonisation.

Pourquoi ce problème résiste

Le problème n'est pas l'absence de volonté de simplifier. L'exploitant nucléaire français conduit des projets de simplification depuis plus de dix ans. Des démarches analogues existent dans les autres secteurs. Mais la complexité croît plus vite que les efforts de simplification ne parviennent à la réduire, parce qu'elle n'est pas uniquement générée en interne.

Le régulateur, pour exercer sa mission de contrôle, a besoin de formalisme : rendre formel ce qui est informel permet l'inspection et la traçabilité. L'ASNR a elle-même reconnu, dans le cadre du Cofsoh[2], qu'elle était une partie prenante dans les différentes sources de complexité. Le législateur national et européen ajoute des exigences sans que les précédentes ne soient systématiquement abrogées. La normalisation internationale impose des cadres de conformité dont la traduction locale alourdit encore le système.

Le mécanisme est donc auto-renforçant et multi-sources. Les travaux conduits à l'ASNR sur les paramètres de la complexité montrent que l'inflation procédurale n'est qu'un symptôme : la complexité tient aussi aux temporalités différentes des parties prenantes, à la connaissance parcellaire des acteurs, aux interfaces mal reliées et aux divergences entre ceux qui conçoivent les règles et ceux qui les appliquent.

Le problème n'est pas seulement le nombre de règles, mais leur type : un référentiel formulé exclusivement en procédures pas à pas produit de la rigidité là où le travail réel exigerait des règles plus flexibles, formulées en compréhension plutôt qu'en extension.

Ce que cela coûte

Les effets documentés vont au-delà du seul allongement des temps de préparation. L'inflation procédurale produit de la perte de sens — les opérateurs cochent des cases sans toujours en comprendre la finalité. Elle produit de la perte de compétence — le temps consacré au *reporting* est pris sur le compagnonnage et la présence terrain. Elle produit une érosion du sens des responsabilités — la multiplication des signatures déresponsabilise plus qu'elle ne protège. Et elle accroît le risque d'erreur par surcharge cognitive. Ces coûts ne sont pas seulement des coûts de sûreté ; ce sont des coûts de disponibilité et de performance industrielle.

---

[1] Lire en particulier le document pédagogique L'analyse des incidents et des accidents : remonter aux causes profondes en recherchant les facteurs organisationnels et humains (FOH), Barpi, octobre 2025.

[2] L'ASN a créé en 2012 le Comité d'orientation sur les facteurs sociaux, organisationnels et humains (Cofsoh) afin de faire progresser la réflexion et les travaux concernant la contribution des personnes et des organisations à la sûreté des installations nucléaires et à la protection des travailleurs.

La dimension informationnelle

Ce problème persiste en partie parce que personne dans l'organisation ne dispose d'une vision consolidée du référentiel dans toute son épaisseur : combien de documents prescriptifs sont en vigueur, lesquels se contredisent, lesquels n'ont jamais été révisés, lesquels ne sont jamais consultés. Mais la dimension informationnelle va plus loin. Les boucles de complexité identifiées dans les travaux du Cofsoh montrent que chaque nouvelle prescription interagit avec les précédentes de manière non linéaire — par les connaissances qu'elle mobilise, les interfaces qu'elle crée, les temporalités qu'elle impose.

Les systèmes d'information industriels actuels — gestion documentaire, GMAO, suites d'asset management ou d'EAM pour industries à actifs critiques, telles que celle d'Hitachi Energy, plateformes QHSE intégrées comme Enablon — ont été principalement conçus pour administrer et tracer des objets : procédures, équipements, ordres de travail, écarts, interventions, contrôles, plans d'action. Leur architecture rend souvent moins lisibles les boucles d'apprentissage qui devraient relier ces objets : du signal faible à l'analyse, de l'analyse à la décision, de la décision à la modification du prescrit, puis du prescrit à sa mise à l'épreuve dans le travail réel.

L'IA entre désormais nativement dans les grandes suites d'asset management, de GMAO, de gestion documentaire ou de QHSE. Mais elle y entre d'abord comme accélérateur de maintenance, de diagnostic, de conformité ou de productivité administrative : aide à la rédaction, détection d'écarts à un référentiel réglementaire externe, analyse documentaire, recommandations d'ordres de travail, assistants conversationnels. Elle améliore la capacité à voir l'état des actifs, des documents ou des écarts ; elle ne garantit pas encore la capacité à comprendre comment l'organisation fabrique, affaiblit ou restaure la cohérence de son propre référentiel[3] de sécurité. Or, ce que la littérature désigne depuis longtemps comme la « prolifération des fonctions de la règle » **[Amalberti 1996]** ou comme l'inflation procédurale réactive **[Dekker 2003]** appelle précisément cette intelligibilité aujourd'hui rarement produite et encore moins exploitée : visibilité des boucles, traçabilité des raisons d'être, repérage des règles devenues redondantes, obsolètes, inapplicables ou contradictoires.

1.2 La préparation est découplée de la maîtrise des risques et de la mémoire opérationnelle

FRAGILITÉ 2

Dans toute industrie à haut risque, les opérations de maintenance, les arrêts programmés et les essais périodiques demandent une préparation rigoureuse. Or, on observe fréquemment que la planification obéit d'abord à une logique de charges, de ressources et d'échéances, et qu'elle traite les risques critiques comme des contraintes additionnelles, parfois trop tard pour reconfigurer réellement le travail.

Le programme START lancé en 2019 par l'exploitant nucléaire français pour améliorer la maîtrise des arrêtés de tranche avec de résultats probants en 2025 formalise une séquence — programmation pluriannuelle, préparation modulaire de qualité, réalisation maîtrisée — dont l'existence même montre que la qualité des arrêts ne dépendait pas seulement de l'exécution.

Le retour d'expérience consolidé par le Barpi et l'IRSN sur l'accidentologie liée à la sous-traitance[4] documente un phénomène structurel : un investissement moindre du donneur d'ordre dans les phases de préparation et dans l'analyse des risques des opérations à faible valeur ajoutée. La rédaction des cahiers des charges, particulièrement lorsque le taux de sous-traitance est élevé et lorsque le rédacteur est éloigné du terrain, tend à se concentrer sur les interventions jugées stratégiques — laissant le flux des opérations courantes (interventions ponctuelles, maintenance programmée, dépannages) sans cadrage écrit comparable. Cette asymétrie entre cadrage des grands chantiers et cadrage du quotidien, documentée pour les phases d'arrêt, se retrouve également en exploitation.

Mais le découplage ne concerne pas seulement les risques. Il concerne aussi la mémoire opérationnelle. Le préparateur qui monte un dossier d'intervention dispose en général de la procédure à jour — c'est le flux le mieux maîtrisé, parce qu'un événement majeur ou une

[3] Constat issu d'une revue des communications publiques 2024-2026 des principaux éditeurs EHS/QHSE intégrés.

[4] Document Barpi/IRSN *Sous-traitance et maîtrise des risques*, décembre 2019.

non-qualité grave finit par modifier la procédure elle-même, et le système d'information permet de s'assurer qu'on utilise le bon indice.

En revanche, il existe une masse considérable de retour d'expérience événementiel positifs et négatifs — en sécurité, en incendie, en environnement, en fiabilité des matériels — qui ne modifie pas les procédures, mais qui serait directement utile à la préparation.

Cette masse de REX est généralement stockée dans des bases de données multiples, chacune dotée de son propre moteur de recherche, de sa propre logique de classement et de ses propres mots-clés. Le préparateur sait que ce REX existe quelque part, mais le trouver suppose de se connecter à plusieurs systèmes, de savoir quoi chercher, et de disposer du temps nécessaire pour trier ce qui est pertinent. En pratique, cette recherche est rarement faite de manière exhaustive.

Ce constat ne renvoie pas à un défaut de conscience professionnelle du préparateur. Il pointe une asymétrie structurelle entre l'exigence implicite — mobiliser le REX disponible — et les moyens disponibles pour le faire : temps alloué, accès aux bases, compétence à chercher dans plusieurs systèmes hétérogènes. La chaîne de supervision qui valide ensuite la préparation hérite d'un dossier dont les sources REX consultées (ou non) ne sont pas un livrable attendu. La traçabilité de la recherche REX n'est ni demandée, ni outillée, ni évaluée. La défaillance du système d'information se transforme ainsi en angle mort partagé de la chaîne préparation-supervision, sans qu'aucun maillon n'en porte explicitement la responsabilité.

La préparation ne se réduit pourtant pas à la consultation du REX. Elle exige aussi de tenir compte du contexte de réalisation : météo prévue, coactivité, disponibilité effective des pièces de rechange, des équipes et des compétences critiques. Cette information-là, contrairement au REX, existe largement dans des systèmes vivants — la GMAO interroge le stock, le planning d'arrêt connaît les coactivités, le SIRH trace les habilitations. La capacité technique d'interroger ces bases est, dans une certaine mesure, déjà acquise : les SI industriels savent appeler d'autres SI.

Ce qui manque n'est donc pas l'accès aux bases, mais l'orchestration de cet accès au service de l'acte de préparation. Le préparateur interroge ces systèmes en série, manuellement, sans qu'aucun mécanisme ne croise pour lui le REX d'une intervention similaire avec la coactivité prévue, la disponibilité des pièces critiques et la compétence présente. La préparation reste ainsi une activité de recollement d'informations dispersées, alors qu'elle gagnerait à devenir un acte de consolidation informée.

Cette fragmentation, qu'il s'agisse du REX historique ou du contexte présent, entraîne une conséquence visible : la préparation est souvent découplée des risques et de la mémoire opérationnelle. L'historique des interventions, l'état réel des équipements, les perturbateurs des arrêts précédents, la disponibilité des compétences critiques existent à des degrés divers dans l'organisation, mais dans des bases distinctes aux logiques incompatibles. L'information qui permettrait de transformer la préparation en acte de maîtrise des risques est là, mais elle n'est ni assemblée ni rendue trouvable au moment où elle serait utile.

Pourquoi ce problème résiste

Le découplage entre planification et maîtrise des risques reflète la manière dont les organisations sont structurées : la planification est portée par une fonction industrielle qui raisonne en termes de délais et de charge ; la maîtrise des risques est portée par une fonction sûreté ou HSE qui intervient en parallèle, souvent en aval. Les situations à haut potentiel de gravité, les perturbateurs du jour et les barrières critiques ne structurent pas toujours comme il le faudrait l'ordonnancement ; ils s'y ajoutent.

La difficulté d'intégrer le REX dans la préparation relève du même mécanisme de découplage. La préparation d'une intervention est alimentée par trois flux de connaissance distincts.

- La procédure à jour — c'est le flux le plus fiable.
- Le REX formalisé, dispersé dans des bases événementielles thématiques — c'est un flux riche, mais difficilement accessible, parce que chaque base constitue un silo.
- Le REX tacite — le savoir de l'agent qui a conduit la même intervention il y a trois ans et qui connaît les particularités de l'équipement, les accès difficiles, les pièges récurrents.

Ce troisième flux ne transite ni par les procédures ni par les bases de données ; il transite par le compagnonnage, les briefings et la mémoire d'équipe. Or c'est précisément ce canal que l'inflation procédurale décrite dans la fragilité précédente érode : le temps consacré au *reporting* et à la gestion documentaire est pris sur la présence terrain et le transfert de savoir entre pairs.

Les industriels travaillent sur ce sujet. Des efforts importants sont consentis pour intégrer le REX dans les préparations futures et pour le pousser vers les intervenants sous une forme exploitable. Mais ces efforts restent fragiles et consomment une énergie importante, parce que le problème de fond n'est pas résolu : il n'existe pas de mécanisme fiable pour mettre en correspondance un enseignement stocké dans une base avec une situation de préparation future.

Le préparateur doit tirer l'information vers lui sans savoir précisément ce qu'il cherche ; l'organisation tente de pousser l'information vers lui sans savoir précisément ce qui est pertinent pour son activité du jour.

Ce que cela coûte

Les effets sur la disponibilité sont directs : allongement des arrêts, reprises de travaux, replanifications tardives, essais répétés, indisponibilités fortuites liées à des défauts de préparation. Lorsqu'un REX pertinent n'a pas été intégré, le risque de reproduire un incident connu augmente — et avec lui le coût de la reprise, de l'investigation et de la perte de confiance des autorités de contrôle.

L'exploitant est confronté à une double peine : un système plus chargé et une mémoire opérationnelle insuffisamment mobilisée pour absorber cette charge.

La dimension informationnelle

La préparation est souvent découplée des risques et de la mémoire opérationnelle en partie parce que les informations nécessaires sont fragmentées entre des systèmes qui ne communiquent pas. L'historique des interventions similaires, l'état réel des équipements, les perturbateurs rencontrés lors des précédents arrêts, la disponibilité des compétences critiques, le REX événementiel multi-domaines — tout cela existe dans l'organisation, mais dans des bases distinctes, avec des logiques de classement incompatibles.

Le préparateur n'a ni une vue consolidée de ce qui s'est passé avant lui sur le même périmètre, ni un mécanisme qui lui signale ce qu'il devrait savoir avant de commencer. L'information qui permettrait de transformer la préparation en acte de maîtrise des risques est là, mais elle n'est ni assemblée ni rendue trouvable au moment où elle serait utile.

1.3 L'entreprise étendue échappe partiellement à la maîtrise

FRAGILITÉ 3

L'exploitation moderne d'une installation à haut risque repose sur un réseau d'entreprises extérieures : prestataires de maintenance, fournisseurs de composants, sous-traitants de rang inférieur, experts détachés. Dans certains secteurs, les effectifs extérieurs représentent la majorité des heures travaillées sur site. La responsabilité de la sécurité industrielle reste formellement du côté de l'exploitant, mais l'exécution est distribuée entre des organisations qui n'ont ni la même culture, ni les mêmes routines, ni les mêmes systèmes d'information.

Dans les ICPE Seveso, les bilans d'inspection montrent que la sous-traitance est souvent gérée par catalogue plutôt que par un processus de maîtrise des risques : absence de listes formalisées de prestataires, processus de sélection flous, implication quasi nulle des salariés extérieurs dans les exercices d'urgence. Dans de nombreuses industries, les sous-traitants peuvent constituer jusqu'à 80% de la main-d'œuvre et restent disproportionnellement exposés aux accidents mortels.

Les industries les plus avancées tentent de construire des réponses numériques à cette fragmentation. Le programme EPR2 a introduit un jumeau numérique partagé dès la conception — une première pour un réacteur nucléaire — et le projet Data4NuclearX, lancé fin 2025 dans le cadre de France 2030, vise à créer un espace souverain de partage de données pour l'ensemble de la filière nucléaire. Mais ces initiatives sont souvent l'apanage de grands projets, conçus dès l'origine avec ces outils. Pour l'immense majorité des installations industrielles

en exploitation, l'entreprise étendue fonctionne encore avec des interfaces documentaires fragmentées, des systèmes d'information qui ne communiquent pas ou peu et des relations structurées davantage par le contrat que par le partage opérationnel.

Pourquoi ce problème résiste

Le problème n'est pas la sous-traitance en soi. C'est la maîtrise d'une entreprise étendue hétérogène alors que la responsabilité de la sûreté reste centralisée du côté de l'exploitant. Une relation principalement structurée par le contrat, le contrôle documentaire et la surveillance a posteriori ne suffit pas à construire une véritable coproduction de la sécurité, une représentation partagée des risques.

Plus le système se fragmente, plus les zones d'ombre augmentent aux interfaces. Les travaux du Cofsoh sur la complexité dans les systèmes à haut risque[5] — ainsi que les réflexions du même comité sur l'articulation entre sûreté réglée et sûreté gérée[6] — montrent que ces interfaces se jouent moins dans la structure organisationnelle que dans la qualité de l'articulation entre le cadrage formel des activités et les capacités d'adaptation collective au réel.

Elles sont aussi des problèmes de temporalités — le rythme de l'exploitant n'est pas celui du prestataire — et de valeurs — la confiance nécessaire au travail partagé se construit sur le terrain, pas dans les clauses d'un contrat. Et plus les modèles de sécurité — le prescrit, le managérial porté, l'opérant — divergent entre donneur d'ordre et sous-traitant, plus les barrières de défense se fragilisent aux points de jonction.

À cette difficulté structurelle s'ajoute désormais un paradoxe. La solution naturelle — partager les données entre donneur d'ordre et sous-traitants pour mieux coordonner le travail — crée un problème que personne n'avait en ne partageant pas. Dès que l'on ouvre des espaces de données communs, la sécurité du système dépend du maillon le plus faible de la chaîne. Le baromètre CESIN 2025 montre que 40% des entreprises françaises identifient des risques critiques liés à leur chaîne de sous-traitance numérique, et qu'un tiers des incidents de cybersécurité sont imputables à des tiers.

Les PME de la filière — sous-traitants de rang 2 ou 3 avec des moyens de cybersécurité limités — deviennent des portes d'entrée vers les systèmes du donneur d'ordre. La directive européenne NIS2, entrée en vigueur en 2024, étend les obligations de cybersécurité aux sous-traitants critiques et passe de 300 à plus de 15 000 entités régulées en France. L'AI Act, le Data Act et le devoir de vigilance convergent dans la même direction : une pression réglementaire croissante sur les interfaces de l'entreprise étendue, qui produit — comme pour la fragilité 1 — des boucles de complexité à l'espace inter organisationnel.

La dimension relationnelle — confiance, visibilité réciproque, capacité à travailler ensemble — doit donc précéder la dimension technologique. Numériser une relation de sous-traitance qui repose encore sur le contrôle documentaire et la méfiance réciproque ne produit pas de l'intégration ; cela produit de la traçabilité sans confiance, et de la donnée sans usage.

Ce que cela coûte

Les conséquences industrielles sont importantes : défauts de qualité, reprises d'intervention, dossiers incomplets, non-conformités, délais de traitement allongés, composants bloqués, indisponibilités prolongées. Une chaîne d'approvisionnement ou une sous-traitance mal intégrée dégrade à la fois la sûreté et la compétitivité de l'exploitation.

Dans les cas graves, une inspection peut suspendre les activités du site. Et le coût de la mise en conformité avec les nouvelles exigences réglementaires de cybersécurité et de traçabilité des données vient s'ajouter à une charge déjà lourde, sans que l'organisation ait encore les outils pour en tirer un bénéfice opérationnel.

---

5 Rapport Cofsoh (ASNR), *Synthèse du cycle de réflexion thématique sur la complexité*, mai 2025. Le Cofsoh y identifie notamment la sur-procéduralisation, les changements permanents et l'ajout de systèmes supplémentaires comme sources de perte de sens et d'effets négatifs sur la maîtrise des risques.

6 Cofsoh GT D (présidé par J. Pariès), *Développer la sécurité — synthèse des travaux du groupe de travail sur l'articulation entre sûreté réglée et sûreté gérée*, ASN.

La dimension informationnelle

L'entreprise étendue souffre d'une fragmentation informationnelle structurelle. Les données de qualification, d'historique d'incidents, de conformité et de performance des prestataires sont dispersées entre les systèmes de l'exploitant, ceux des entreprises extérieures et ceux des organismes de certification. Personne ne dispose d'une vision consolidée du risque fournisseur.

Mais le problème a changé de nature. Il ne s'agit plus seulement de relier des données dispersées : il s'agit de les relier dans des conditions de sécurité, de souveraineté et de gouvernance que les infrastructures actuelles ne fournissent pas. Les projets comme Data4NuclearX montrent que la filière a identifié l'enjeu. Leur calendrier — déploiement opérationnel prévu à partir de 2028 — montre aussi la distance qui sépare la prise de conscience de la capacité opérationnelle.

1.4 Les signaux faibles ne remontent pas suffisamment — ou ne sont pas assemblés

FRAGILITÉ 4

La vigilance sur les signaux faibles[7] est un principe fondateur des organisations à haute fiabilité. Au-delà même des quasi-incidents — qui doivent en théorie être analysés comme des accidents potentiels — la détection des précurseurs reste fragile dans tous les secteurs, et pour des raisons qui vont au-delà de la seule culture de signalement.

Les signaux existent, mais ils proviennent de sources hétérogènes qui ne se parlent pas toujours.

- ▷ Le canal le plus visible est le déclaratif humain : signalements, quasi-accidents, observations terrain, mais aussi éléments dans les comptes rendu, mains courantes, cahier de quart…
- ▷ Mais les systèmes de supervision instrumentée produisent eux aussi des signaux — dérives de paramètres, sollicitations anormales des sécurités automatiques, alarmes récurrentes — qui ne sont pas ou rarement classés comme des « événements » et qui restent dans les bases de données techniques sans qu'un rapprochement avec les signalements humains soit tenté.
- ▷ Les données opérationnelles (GMAO, temps d'intervention, taux de défaillance…) portent d'autres indices de fragilité qui, eux non plus, ne sont pas toujours reliés au flux de détection des précurseurs.

Chaque source a sa propre logique, sa propre temporalité, son propre système d'information — et souvent son propre propriétaire organisationnel.

Dans les ICPE Seveso, l'Ineris rappelle[8] que la plupart des accidents majeurs sont précédés de signaux faibles et de précurseurs, qui restent perçus, mais banalisés par un mécanisme de normalisation de la déviance — autrement dit, qui ne deviennent évidents qu'après la survenue de l'accident.

Pourquoi ce problème résiste : Trois mécanismes se conjuguent

Le premier est **architectural**. Le système de détection repose trop souvent sur un canal unique — le déclaratif spontané — et sur des catégories déjà stabilisées. Ce qui ne rentre pas dans les cases n'est pas vu. Les critères de triage entre le bruit quotidien (faible potentiel de gravité et d'apprentissage) et le signal critique (haut potentiel de gravité et d'apprentissage…) restent largement implicites, peuvent dépendre du contexte, ce qui les rend difficiles à appréhender. Les événements récupérés sans conséquence — ceux qui sont parfois les plus riches en enseignements — ne déclenchent pas d'analyse approfondie parce qu'ils n'ont « rien produit ». Parallèlement, la multiplication des capteurs et des alarmes automatiques produit son propre problème : la saturation. Plus on capte, plus on noie le signal dans le bruit — sauf si un mécanisme de triage intelligent existe entre la captation et l'interprétation humaine.

Le deuxième est le **silence organisationnel** amplifié aux interfaces de l'entreprise étendue. Les salariés des entreprises extérieures — souvent les plus proches des situations à risque —

---

7 Un signal faible se définit, dans le contexte de la sécurité industrielle, comme une information ambiguë, partielle ou dispersée qui — si elle est correctement détectée et interprétée — peut révéler un risque latent en amont de sa concrétisation. Il se distingue du presqu'accident, événement avéré et identifiable, par son caractère diffus et facilement banalisé. La détection des signaux faibles est l'un des marqueurs des organisations à haute fiabilité.

8 Document *Presque accidents et Risque d'accident majeur état de l'art* (DRA-37), 2004.

sont structurellement moins enclins à signaler un écart à leur client. L'asymétrie de pouvoir contractuelle, la crainte de conséquences commerciales et l'absence de protection du signalement inter-entreprises créent un filtre supplémentaire entre le signal et sa remontée. **[Morrison et Milliken 2000]** ont montré que le silence organisationnel est un phénomène collectif, produit par le système, pas par la lâcheté individuelle.

Les travaux de **[Rocha 2014]** sur le silence des sous-traitants dans les industries à risque françaises confirment que la frontière entre organisations agit comme un filtre qui prive le système de ses capteurs les plus précieux. La fragmentation décrite dans la fragilité 3 a donc un effet direct sur la détection.

Le troisième est **la superficialité des analyses qui suivent le triage**. Même lorsqu'un signal remonte et qu'un événement est classé HIPO, l'analyse s'arrête trop souvent aux causes apparentes — défaillance technique, écart de procédure — sans explorer les causes organisationnelles et systémiques. Un signal capté puis mal interprété est un signal perdu. Ce que cela produit — des analyses par événement et non par type, une incapacité à identifier les motifs récurrents et à questionner le modèle de sécurité — sera développé dans la fragilité suivante.

Ce que cela coûte

Sans détection précoce, l'installation est exposée à la **surprise organisationnelle**. Ce qui n'a pas été vu à temps réapparaît sous forme d'incident, d'arrêt imprévu, de campagne de contrôles correctifs ou de reprise de travaux. La disponibilité souffre alors d'un déficit d'anticipation, pas seulement d'un défaut de réaction. Certaines dérives — un paramètre qui glisse lentement, un temps de réponse qui s'allonge, un taux de sollicitation des sécurités qui augmente — ne sont classées comme « problème » par personne tant qu'elles n'ont pas produit de conséquence visible. C'est pourtant dans ces dérives silencieuses que se prépare l'érosion des marges[9].

La dimension informationnelle

Les signaux faibles ne sont pas absents : ils sont dispersés et parfois incompatibles. Le signalement humain parle en langage naturel dans une ou des bases de données (sûreté, sécurité, incendie, environnement…). Le capteur parle en séries temporelles dans un historien de données. La GMAO parle en codes d'intervention. L'observation managériale parle en compte rendu de visite.

Lorsqu'ils existent, les mécanismes qui les rapprochent, les mettent en correspondance sémantique ou les corrèlent avec les scénarios à haut potentiel de gravité reposent largement sur l'effort individuel et l'expérience tacite — rarement sur une orchestration outillée. Chaque source vit dans son silo, avec son vocabulaire, sa temporalité et son propriétaire métier.

L'information qui permettrait de voir une dérive systémique — en reliant une augmentation des sollicitations d'une sécurité automatique, un signalement terrain sur la même zone et un historique de maintenance dégradé sur le même équipement — est là, dans l'organisation, mais elle n'est ni assemblée ni rendue lisible.

1.5 Le retour d'expérience ne transforme pas assez — et ne questionne pas le modèle de sécurité

FRAGILITÉ 5

Les industries à haut risque disposent de systèmes de retour d'expérience structurés, parfois très riches en données. Le problème n'est pas l'absence de REX au sens du dispositif de collecte et d'analyse : c'est que le REX ne se referme pas toujours sur une transformation effective et durable des pratiques et des barrières. L'inspecteur général pour la sûreté nucléaire d'EDF a employé en 2025 une formule éclairante : le « retour d'expérience en boucle ouverte ». On collecte, on analyse, on prescrit des actions correctives — mais la boucle ne se referme pas sur une modification réelle et durable des conditions de travail. L'ASNR a noté en 2025 que certains sujets identifiés depuis plusieurs années en radioprotection ne montrent pas de tendance nette à l'amélioration, malgré un volume important d'analyses produites.

[9] Ensemble des écarts disponibles — techniques, temporels, cognitifs et collectifs — qui permettent à une installation et à ses équipes d'absorber un aléa sans basculer dans l'accident. Les accidents graves ne résultent pas du franchissement d'une limite isolée, mais de l'érosion silencieuse et combinée de plusieurs marges (Amalberti, Rasmussen).

Un phénomène aggrave cette situation : les analyses restent presque toujours au niveau de l'événement individuel. **On analyse souvent l'accident, pas assez souvent le type d'accident**. On traite le cas, pas le mécanisme. Les organisations capitalisent des dizaines, parfois des centaines d'analyses, mais ne les croisent pas toujours pour identifier les motifs récurrents, les familles de scénarios et les fragilités structurelles qui les produisent.

Un second phénomène se superpose : pris individuellement, les supports de REX plafonnent très souvent au niveau des conditions locales (pourquoi l'opérateur a fait ceci, pourquoi la barrière n'a pas tenu) sans atteindre le niveau des mécanismes organisationnels (pourquoi l'organisation avait accepté cette configuration) ni celui de la gouvernance (quelle doctrine l'autorise durablement). Ce plafond n'est pas un défaut de compétence des analystes ; c'est un effet de système — format de restitution, habitudes, absence de standard explicite de profondeur. Personne ne pose alors la question de niveau 2 : que nous apprennent nos vingt derniers HIPO, pris ensemble, sur la solidité de notre modèle de sécurité ?

Pourquoi ce problème résiste

L'apprentissage organisationnel à partir des événements et des situations à haut potentiel peut se déployer à trois niveaux de profondeur — et la plupart des organisations peinent encore à systématiser les deux derniers.

- **Premier niveau : corriger l'écart**. On analyse l'événement, on identifie ce qui n'a pas fonctionné, on corrige localement la barrière défaillante, on rappelle la règle. C'est la forme la plus répandue, et la plupart des systèmes de REX la pratiquent.
- **Deuxième niveau : remonter aux causes systémiques**. On dépasse la correction de l'écart pour conduire une analyse approfondie sur l'événement — couvrir les axes pertinents (technique, humain, organisationnel, culturel), remonter au-delà des conditions locales jusqu'aux mécanismes organisationnels qui ont rendu l'événement possible et en rendront d'autres possibles tant qu'ils ne seront pas instruits. On corrige, on traite, on modifie les barrières et les conditions de travail, mais sans questionner les hypothèses qui définissent ce qu'on considère comme une cause acceptable et comme une barrière pertinente.
- **Troisième niveau : questionner le modèle de sécurité**. On dépasse l'analyse cas par cas pour une relecture transversale qui interroge les hypothèses qui fondent la réponse à la question *« qu'est-ce qui fait qu'il n'y aura pas d'accident grave dans notre organisation ? »*. Ces hypothèses sont-elles encore valides ? Quels patterns récurrents traversent nos événements et désignent une vulnérabilité systémique que les analyses unitaires n'avaient pas vue ? C'est ce que Argyris et Schön nomment *double-loop learning* — et il est rarement pratiqué.

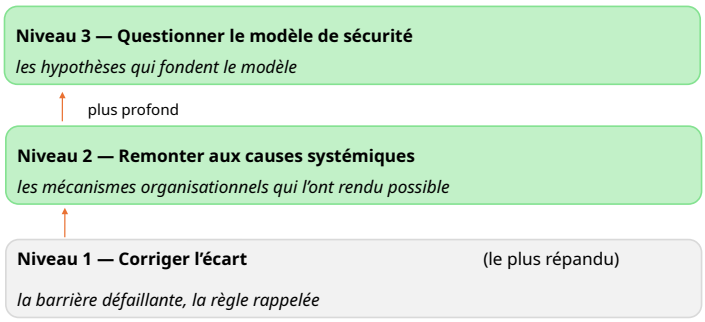

Fig. 1.1 Trois niveaux de profondeur de l'apprentissage organisationnel. La majorité des organisations s'arrêtent souvent au niveau 1.

[Drupsteen et Guldenmund 2014] ont montré que la littérature sur l'apprentissage par les incidents se concentre massivement sur les étapes amont — collecte, analyse, identification des causes — et néglige les étapes aval : le stockage, le partage, la réutilisation effective et la vérification d'efficacité. Carroll (MIT), étudiant les centrales nucléaires américaines, observe que les managers apprécient les rapports qui proposent des actions correctives logiques, tandis que les analystes valorisent les rapports qui découvrent des enseignements transférables à d'autres contextes. Ce décalage produit des plans d'action locaux qui rassurent sans toucher le système **[Carroll 1998]**.

Il y a plus : le retour d'expérience lui-même est un producteur de complexité. Les travaux conduits dans le cadre du Cofsoh montrent que chaque analyse débouche sur de nouvelles prescriptions, de nouveaux modes opératoires, de nouveaux documents de preuve, qui s'ajoutent à un référentiel déjà saturé. Le dispositif censé améliorer le système alimente la spirale décrite dans la fragilité 1.

Enfin, les organisations disposent de données structurées considérables — résultats de formations dans le SIRH, scores de questionnaires dans le LMS, historiques de maintenance dans la GMAO, données de planification — qui pourraient enrichir les analyses individuelles et surtout faire émerger des tendances de niveau 2. Mais ces données vivent dans des silos séparés, avec des propriétaires différents, et ne sont presque jamais croisées avec les données d'événements.

Point d'attention

Croiser les données RH individuelles (ancienneté, parcours, résultats de formation) avec les données d'événements ressemble à un gisement d'apprentissage. C'en est un piège dès que la culture juste n'est pas tenue : perçus comme un outil de mise en cause, ces croisements renforcent le silence organisationnel au lieu de le réduire — et tarissent la matière même qu'ils prétendaient exploiter. S'y ajoute le risque de fausses corrélations : observer que les intérimaires déclarent plus d'accidents ne dit rien tant qu'on n'a pas séparé l'effet de la formation de celui de l'exposition aux tâches les plus à risque.

Ce que cela coûte

Le coût est massif, mais diffus : répétition de défaillances proches, campagnes de rattrapage, accumulation de plans d'action non suivis d'effet, temps d'encadrement consommé dans la correction de symptômes, pertes de production récurrentes sans apprentissage consolidé. Chaque incident redondant gaspille des ressources et diminue la confiance des autorités de contrôle.

Et le REX qui ne redescend pas dans la préparation des activités futures — point développé dans la fragilité 2 — prive le préparateur de l'ingrédient qui lui serait le plus utile.

La dimension informationnelle

Le REX ne transforme pas assez parce que les enseignements sont archivés par événement, pas par mécanisme. Les récurrences ne sont pas détectées parce que les analyses sont classées par date et par lieu, pas par type de scénario, par barrière défaillante ou par perturbateur récurrent.

Mais le problème va au-delà de l'archivage des analyses elles-mêmes. L'organisation dispose de multiples couches de données — événementielles, opérationnelles, planification, RH, formation, supervision instrumentée… — qui vivent dans des systèmes séparés et dont le croisement permettrait de faire émerger des tendances que personne ne voit aujourd'hui : corrélation entre ancienneté de formation et récurrence d'écarts, entre charge de travail et densité d'événements, entre taux de couverture des habilitations et fréquence des HIPO. Ces rapprochements sont aujourd'hui difficiles à réaliser de manière systématique — non pas parce que les données n'existent pas, mais parce qu'aucun système ne les relie et qu'aucun processus organisationnel ne les demande.

1.6 Les indicateurs rendent visible l'activité plus que la robustesse du système

FRAGILITÉ 6

Le pilotage des installations à risque s'appuie sur des indicateurs qui, pour la plupart, mesurent l'activité réalisée plutôt que la santé réelle des défenses. Le management se focalise souvent sur ce qui est mesurable et consolidable rapidement : délais d'arrêt, volumes de maintenance, nombre d'événements, taux de conformité. En sécurité au travail, la baisse du taux de fréquence peut masquer une dégradation des barrières de sûreté des procédés — c'est l'une des leçons majeures de l'accident de Texas City (2005).

Dans le pétrole, l'IOGP a conduit un travail de fond pour distinguer les indicateurs de sécurité personnelle de ceux de sécurité des procédés, et le taux d'accidents avec arrêt a été jugé insuffisant pour piloter la prévention des événements majeurs.

Le même biais affecte le pilotage de l'entreprise étendue. Les indicateurs de sous-traitance sont souvent contractuels et administratifs — conformité des habilitations, taux de réalisation des plans de prévention… — et ne rendent pas toujours compte de la qualité réelle de l'intégration opérationnelle. Le taux de fréquence des entreprises extérieures, peut coexister avec une sous-déclaration structurelle liée au silence organisationnel décrit dans la fragilité 4. Un indicateur « vert » chez le sous-traitant peut masquer une dégradation invisible des interfaces.

Pourquoi ce problème résiste

Les indicateurs ne décrivent pas seulement le système ; ils orientent les comportements. Si les métriques dominantes valorisent la clôture administrative, la conformité de façade ou le respect apparent du planning, les collectifs apprennent à optimiser ces résultats — même lorsque la robustesse réelle des barrières est plus faible. Un système peut ainsi devenir plus gouverné et moins sûr, plus piloté et moins apprenant.

Mais le biais d'indicateurs n'est pas seulement un problème de choix managérial. C'est aussi une conséquence de la fragmentation informationnelle décrite dans les fragilités précédentes. On pilote avec ce qu'on sait produire facilement — les données volumétriques, les taux, les délais — parce que les données qui rendraient visible la robustesse réelle exigent un travail de consolidation que l'organisation n'a très souvent ni le temps ni les outils de conduire.

Construire un indicateur de santé des barrières critiques suppose de croiser les données de supervision (sollicitations des sécurités automatiques, dérives de paramètres), de maintenance (disponibilité des équipements, reports d'intervention), de REX (récurrence des défaillances par type) et d'observation terrain — précisément les sources que les fragilités 4 et 5 ont montré disjointes. Tant que ce croisement n'est pas fait, les indicateurs disponibles reflètent l'effort consenti, pas le résultat obtenu.

Il y a un piège supplémentaire : la tentation de « refondre les indicateurs » peut produire exactement l'effet inflationniste décrit dans la fragilité 1. Chaque nouvel indicateur demande un processus de collecte, de validation, de *reporting* et de revue qui s'ajoute à la charge. Si la refonte n'est pas disciplinée — peu d'indicateurs, mais alignés sur les mécanismes qui comptent, et effectivement utilisés dans les arbitrages — elle alimente la spirale de complexité au lieu de la réduire.

Ce que cela coûte

Des indicateurs mal alignés produisent des décisions sous-optimales : on protège des objectifs visibles à court terme en fragilisant les conditions de disponibilité future. Les coûts apparaissent plus tard — retards, reprises, défauts, incidents répétés, surcharge des équipes, perte de flexibilité. À l'inverse, des indicateurs robustes permettent d'anticiper la maintenance, de mesurer la performance réelle et de prioriser les investissements sur les barrières qui comptent.

La dimension informationnelle

C'est ici que convergent les fils des cinq fragilités précédentes. Les données nécessaires au pilotage de la robustesse — taux de disponibilité des barrières critiques, délai de fermeture effective des actions REX, corrélation entre perturbateurs et événements, couverture réelle des compétences critiques — existent dans l'organisation, mais dans les mêmes silos que ceux décrits tout au long de ce chapitre.

Le système HSE porte les événements. La GMAO porte la maintenance. Le SCADA porte les paramètres de supervision. Le SIRH porte les compétences et les formations. Les bases de REX portent les analyses. Ces systèmes dialoguent peu entre eux, et les passerelles existantes sont généralement conçues pour le *reporting*, non pour le pilotage de la robustesse. Il manque un mécanisme qui agrège ces sources en une vision consolidée — disponibilité réelle des barrières, fermeture effective des actions REX, corrélation perturbateurs/événements, couverture des compétences critiques — orientée vers l'anticipation plutôt que vers la conformité.

Tant que cette consolidation n'est pas faite, le pilotage restera captif des indicateurs de surface — non par choix délibéré, mais largement par défaut d'accès consolidé aux données qui permettraient de faire autrement.

Une architecture cohérente, pas un système qui ne fonctionne pas

Les six fragilités décrites dans ce chapitre forment une architecture « cohérente » où chaque difficulté alimente les autres. Les trois premières — complexité du référentiel, fragilité de la préparation, opacité de l'entreprise étendue — jouent un rôle de causes amont : elles façonnent les conditions dans lesquelles le travail est préparé, organisé et exécuté. Cette interaction relève d'une spirale de complexité : chaque fragilité produit, par les dispositifs censés la corriger, de la complexité supplémentaire qui affecte les autres. Les deux suivantes — détection des signaux faibles, bouclage du retour d'expérience — jouent un rôle de révélateurs : elles déterminent ce que le système voit de lui-même et ce qu'il en fait. La sixième — alignement des indicateurs — structure les arbitrages et peut renforcer ou atténuer les cinq autres.

Il serait erroné de lire cette architecture comme le portrait d'un système qui ne fonctionne pas. Les industries à haut risque sont précisément celles qui ont le plus investi dans la maîtrise de ces fragilités — et les résultats sont mesurables : les taux d'accidents graves ont diminué dans la plupart des secteurs. Ce qui est en jeu n'est pas l'existence des dispositifs de sûreté — ils sont robustes — mais la difficulté à maintenir leur cohérence opérationnelle dans un environnement qui se complexifie structurellement. C'est sur cette difficulté précise que les promesses de l'IA méritent d'être interrogées.

Deux mécanismes transversaux aggravent l'ensemble :

- **L'amnésie organisationnelle** : La rotation des managers — souvent voulue comme politique de développement des compétences — fait perdre la mémoire des arbitrages, des compromis et des raisons qui ont fondé les choix passés. L'AIEA a formalisé ce risque comme un enjeu majeur pour les organisations nucléaires, mais le phénomène dépasse largement ce secteur.

  Dans l'entreprise étendue, il prend une forme particulière : quand un chef de chantier sous-traitant ou un responsable contrat côté donneur d'ordre change, c'est la relation de confiance construite sur le terrain qui disparaît — une mémoire relationnelle qu'aucun système d'information ne porte. Il y a là un paradoxe que les grandes organisations doivent affronter : un directeur de site qui reste trois ans n'a pas le temps de voir les effets de ses décisions sur la sûreté ; les conséquences d'un arbitrage apparaissent souvent bien plus tard.

- **Le tarissement de la matière apprenante** : Un mécanisme transversal, propre à l'ère de l'IA, traverse l'ensemble. Quand l'IA s'entraîne sur le matériau produit par l'effort cognitif humain — REX, analyses, observations terrain — et qu'elle réduit cet effort en le substituant, elle érode progressivement la matière qu'elle exploite (*cf.* la figure 1.2). **[Acemoglu et al. 2026]** ont formalisé cette dynamique : sous certaines conditions, le système peut basculer vers un équilibre où la connaissance générale s'évapore en dépit d'une assistance personnalisée toujours plus précise.

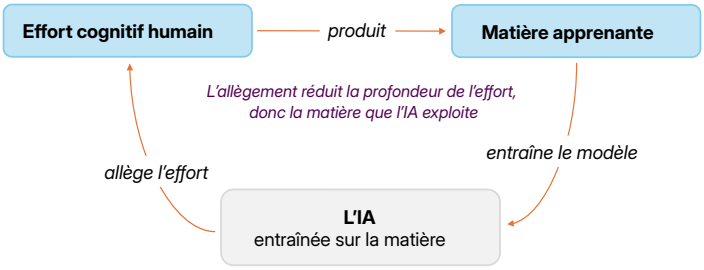

Fig. 1.2 Le mécanisme de tarissement de la matière apprenante.

Les fragilités 4 (signaux faibles) et 5 (bouclage du REX) sont les zones les plus exposées, parce que leur production dépend d'une qualité d'attention humaine qui est élastique et difficilement mesurable.

Ces deux mécanismes — amnésie organisationnelle et tarissement de la matière apprenante — montrent que le problème n'est pas un manque de dispositifs, mais une difficulté à maintenir la cohérence d'un système qui se complexifie et qui oublie plus vite qu'il n'apprend.

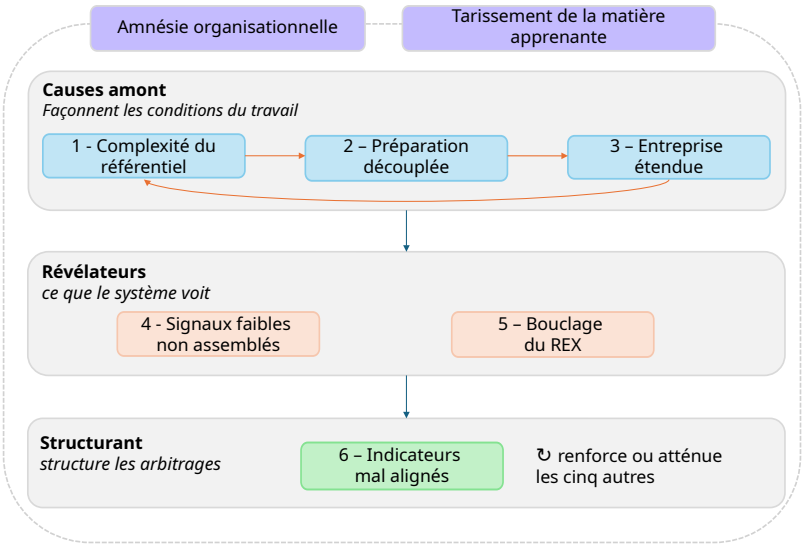

Fig. 1.3 *L'architecture des six fragilités. Deux mécanismes transversaux aggravent l'ensemble : l'amnésie organisationnelle et le tarissement de la matière apprenante.*

Un périmètre informationnel assumé

Les six fragilités décrites partagent une caractéristique : elles comportent toutes une dimension informationnelle qui explique en partie leur persistance. Information dispersée dans des silos qui ne communiquent pas suffisamment. Information pas ou peu corrélée entre des sources hétérogènes — déclaratif humain, supervision instrumentée, données opérationnelles, données RH et formation. Information enfouie dans des archives classées par événement et non par mécanisme. Information tacite portée par des personnes et perdue à chaque départ.

Cette fragmentation n'est pas un problème technique secondaire : c'est une cause structurelle de la persistance des six fragilités. C'est ce choix de lecture qui définit le périmètre du présent document — non pour affirmer que l'IA résout ces problèmes, mais pour examiner ce qu'elle peut rendre visible, relier ou accélérer, à condition que le système sous-jacent soit suffisamment sain pour en tirer profit.

Ce périmètre laisse hors champ un autre registre de la sécurité industrielle — l'exposition physique de l'humain à des situations dangereuses — qui appelle des réponses d'une autre nature : robotique d'inspection en zones hostiles, drones en milieux confinés ou irradiés. L'IA y joue un rôle documenté, et l'apport en sécurité y est souvent le plus immédiatement démontrable parce qu'il soustrait physiquement l'humain au danger.

Les paragraphes qui précèdent ont rappelé que ces fragilités sont à géométrie variable — leur saillance n'est pas la même dans une centrale nucléaire, un site chimique Seveso, une plateforme pétrolière ou une usine de l'industrie chimique. C'est au lecteur d'évaluer, pour son propre contexte, lesquelles sont les plus critiques et lesquelles interagissent le plus fortement.

Question ouverte

Dans un monde où la complexité et l'incertitude croissent structurellement — où il s'agit d'apprendre à « vivre avec » plutôt qu'à tout contrôler — que se passe-t-il quand on introduit l'intelligence artificielle dans un système dont les fondations sont encore fragiles ?

1.6.1 Une clé de lecture pour la suite

C'est le régime d'usage — et non la technologie elle-même — qui détermine les effets sur la compétence, la vigilance, la responsabilité et le modèle de sécurité. Une même technologie LLM peut, selon le régime d'usage qui lui est donné, renforcer le jugement humain ou l'éroder, rendre visible une fragilité ou la masquer, faciliter l'apprentissage ou court-circuiter le débat. Cette idée — qu'il faut qualifier l'usage avant la technologie — est l'une des clés de lecture du document. Elle prendra sa forme opératoire au chapitre 3, avec la grille *déléguer, assister, sanctuariser.*

Point clé

L'enjeu n'est plus seulement de savoir si l'IA est embarquée dans une caméra, une GMAO ou un outil QHSE. L'enjeu est de comprendre que l'IA devient progressivement une couche d'architecture industrielle : elle relie les données, interprète les signaux, assiste les décisions et produit une représentation du système. Dans les industries à haut risque, cette représentation ne peut pas être laissée au seul modèle de données. Elle doit être gouvernée par un modèle de sécurité explicite.

2

Ce que l'intelligence artificielle sait faire — et ce qu'elle ne sait pas faire

Le chapitre précédent a posé un diagnostic : six fragilités structurelles traversent les industries à haut risque et résistent aux solutions classiques. Chacune comporte une dimension informationnelle — information dispersée, non corrélée, enfouie — qui est l'une des raisons de sa persistance. C'est cette dimension qui justifie d'examiner ce que les systèmes à base d'IA peuvent apporter.

L'introduction a posé que ce Cahier parle d'une mosaïque hétérogène plutôt que d'une « IA » monolithique. Ce chapitre approfondit la distinction et la rend opérationnelle pour la suite. Avant de présenter les familles techniques mobilisables, deux filtres successifs sont nécessaires pour délimiter ce dont on parle :

- Premier filtre : **informationnel ou matériel** : Les systèmes IA déployés dans l'industrie agissent de deux manières très différentes. Les uns produisent de l'information qu'un humain exploite : détection d'anomalies, reconnaissance visuelle, synthèse de corpus, simulation prospective. Les autres agissent matériellement à la place de l'intervenant ou en lieu et place de son jugement : robotique autonome, télé-opération, exosquelettes, dispositifs d'alerte autoritaire qui déclenchent une action sans passer par une décision humaine. Nous nous focalisons dans ce document sur les premiers — les systèmes IA informationnels, dont la finalité est de nourrir une décision humaine, y compris quand cette finalité est la protection de l'humain sur un chantier (vision par ordinateur appliquée à la supervision, par exemple).
- Second filtre : **IA-outil ou IA-système porteur d'un modèle de sécurité**. Parmi les systèmes IA informationnels, la massivité du déploiement actuel se trouve dans une famille particulière : des IA spécialisées, embarquées dans des outils ou des équipements, conçues et entraînées par des fournisseurs, vendues clés en main avec leur usage prescrit. Caméras intelligentes pour la vidéosurveillance d'intrusion, algorithmes de reconnaissance de défauts intégrés aux contrôles non destructifs, capteurs autonomes de détection d'événement.

  Ces systèmes ne portent pas, en eux-mêmes, un modèle de sécurité ; ils portent un modèle de reconnaissance ou de détection. Leur gouvernance relève de logiques classiques — qualification fournisseur, validation métrologique, assurance qualité, doctrine de sûreté de site — qui préexistent à l'IA et que l'IA ne reconfigure pas en profondeur. Ce document ne les traite pas pour eux-mêmes, sans nier leur criticité possible : une IA d'analyse CND sur le circuit primaire d'une centrale contribue in fine à la sûreté de l'installation, et ses faux négatifs ont les mêmes conséquences potentielles qu'une inspection humaine défaillante. Le chapitre consacré au déploiement reviendra sur ce qu'elles signifient comme contexte préalable pour le donneur d'ordre.

L'objet propre de ce *Cahier* est celui des systèmes IA qui interviennent, ou prétendent intervenir, dans la résolution des fragilités identifiées au chapitre 1 : intégration du REX dans la préparation, analyse de signaux faibles dispersés, génération assistée de règles, anticipation d'événements à haut potentiel, qualification des marges de manœuvre, présence terrain augmentée. Ces systèmes-là, qu'il s'agisse de grands modèles de langage, de plateformes d'analyse de données métier ou de jumeaux numériques contextualisés, portent (ou risquent de porter, selon la manière dont ils sont déployés) un modèle de sécurité. Ce qu'ils donnent à voir, ce qu'ils dissimulent, ce qu'ils prescrivent ou suggèrent configure progressivement la culture de sécurité de l'organisation qui les utilise. C'est cette catégorie qui justifie un examen détaillé.

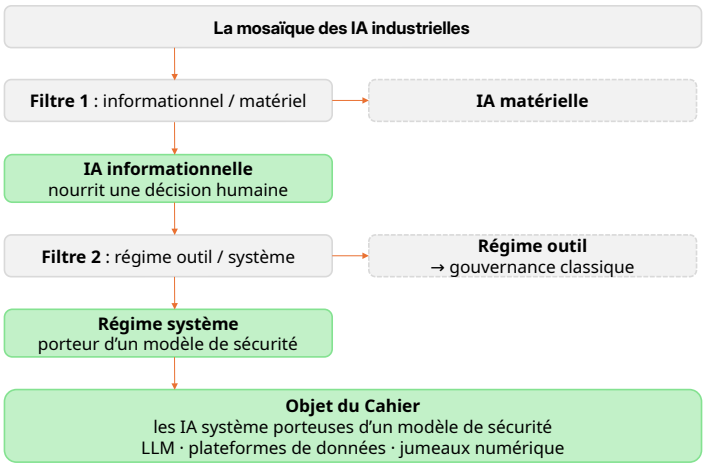

Fig. 2.1 Chemin retenu pour ce document, qui s'intéresse aux systèmes d'IA informationnels dont la finalité est d'alimenter une décision humaine, et qui intègrent de façon implicite une manière de modéliser la sécurité.

Une précision sur les capacités techniques : toutes les technologies décrites dans ce chapitre sont des IA à domaine d'application étroit — des systèmes conçus pour des tâches spécifiques. L'intelligence artificielle générale n'existe pas en contexte industriel, et les systèmes les plus avancés (grands modèles de langage compris) restent des spécialistes de tâches définies, qui peuvent enchaîner des opérations sans pour autant former une intelligence unifiée. En vingt ans d'évolution de l'analyse des données industrielles, le facteur limitant n'a jamais été la technologie. Il a toujours été organisationnel : la qualité des données, le cloisonnement des systèmes, le fossé entre spécialistes de la donnée et experts métier, et surtout la capacité à transformer un résultat d'analyse en action opérationnelle.

2.1 Les familles d'IA informationnelles : ce que le praticien doit en savoir

Sept familles technologiques composent aujourd'hui l'offre d'IA mobilisable en industrie à risque. Toutes sont des *IA à domaine étroit*[1] — des outils conçus pour des tâches spécifiques, pas des intelligences générales. La présentation qui suit est délibérément ramassée : pour chaque famille, on indique ce qu'elle rend lisible au praticien, l'irritant qu'elle est susceptible de traiter, et le risque principal qu'elle introduit.

Machine Learning supervisé et non supervisé

Famille de modèles statistiques qui apprennent des règles de décision à partir de données historiques.

- **Pour le praticien**, c'est ce qui rend exploitables les données de capteurs accumulées depuis des années sans usage analytique : signatures vibratoires, dérives thermiques, profils de consommation.
- **L'irritant traité** est celui de l'équipement qui défaille « sans prévenir » alors que les signaux annonciateurs étaient présents dans les données.
- **Le risque associé** est la *dérive silencieuse* : un modèle entraîné sur la configuration d'hier se dégrade sans le signaler quand l'installation, le mode opératoire ou la météo changent — et les *cygnes noirs*, scénarios jamais rencontrés, échappent par construction à l'apprentissage.

[1] IA à domaine étroit : système entraîné et qualifié pour une tâche spécifique et bornée — détection d'anomalie, classification d'événements, prévision d'usure d'un composant —, par opposition aux IA généralistes qui prétendent traiter un large éventail de tâches.

2.1.1 Deep learning (réseaux de neurones profonds)

Famille qui repose sur des réseaux de neurones à plusieurs couches, capables d'extraire automatiquement des motifs dans des données massives ou non structurées.

- ▷ **Pour le praticien**, c'est ce qui permet de traiter ce que le ML classique ne sait pas traiter : images complexes, séries temporelles longues et bruitées, signaux multimodaux.
- ▷ **L'irritant traité** est celui des phénomènes dont la signature ne tient pas dans une variable simple — corrosion par piqûres sur une image d'inspection, défaut de roulement noyé dans un spectre vibratoire…
- ▷ **Le risque associé** est l'opacité : le réseau produit une réponse sans que l'on puisse expliquer le raisonnement qui y conduit, ce qui pose un problème direct de démonstrabilité dans les contextes où l'autorité de contrôle exige la traçabilité du jugement.

2.1.2 Vision par ordinateur

Famille spécialisée dans l'extraction d'information à partir de flux vidéo et d'images, fondée principalement sur du **deep learning** convolutif. C'est, parmi les sept familles présentées ici, celle où la distinction entre IA-outil et IA-système porteur d'un modèle de sécurité, posée en ouverture du chapitre, est la plus immédiatement lisible.

- ▷ En **IA-outil**, la même technologie sous-jacente recouvre des usages très différents — vidéo-surveillance d'intrusion, comptage de personnes, reconnaissance de défauts en contrôle non destructif, analyse d'images d'inspection. Ces dispositifs ne portent pas un modèle de sécurité (ils portent un modèle de reconnaissance), même quand leurs sorties contribuent *in fine* à la sûreté de l'installation. Leur gouvernance relève de logiques propres — qualification fournisseur, validation métrologique, sûreté de site, RGPD.
- ▷ En **IA-système** — celui que ce *Cahier* traite — la vision par ordinateur est intégrée à une architecture sociotechnique de supervision : observation de chantier à grande échelle (port d'EPI adapté au risque, franchissement de zone interdite, posture à risque, fuite visible, écart gestuel…), avec une finalité d'apprentissage collectif et une intégration des constats au REX. Ce qu'elle rend lisible au praticien, dans ce type d'IA, c'est ce que l'observation terrain ne peut pas être partout, tout le temps, et dont les constats restent souvent oraux et non capitalisés.
- ▷ **L'irritant traité** est celui de la perte d'information sur ce qui se passe réellement en activité, par défaut de présence et défaut de capitalisation.
- ▷ **Le risque associé** est double : *fatigue d'alerte* si les faux positifs s'accumulent, et surtout glissement de la *sécurité* vers la *surveillance* si le dispositif n'est pas cadré socialement — auquel cas il détruit la confiance qui est le pilier de la culture de signalement, et fait basculer le système, en pratique, hors des IA-système dont parle ce *Cahier*.

2.1.3 NLP et grands modèles de langage (LLM)

Famille de modèles qui traitent du texte non structuré — de l'exploitation de texte classique aux LLM génératifs en passant par les approches d'augmentation par récupération (RAG).

- ▷ **Pour le praticien**, c'est ce qui rend lisibles à l'échelle les corpus jusqu'ici dormants : milliers de REX, comptes rendus d'audit, Cahiers de quart, commentaires libres dans les ordres de travail…
- ▷ **L'irritant traité** est celui de la connaissance enfouie dans des récits qu'aucun individu seul ne peut lire en entier — et qui sont précisément les données les plus riches pour comprendre le « pourquoi » des événements.
- ▷ **Le risque associé** est l'*hallucination* : la production d'une synthèse plausible, mais fausse, présentée avec la même assurance qu'une réponse correcte. Risque accentué sans prise en compte du modèle de sécurité.

2.1.4 Jumeaux numériques

Famille hybride combinant simulation physique, ML et flux de données capteurs, qui maintient en parallèle de l'installation réelle une réplique numérique alimentée en continu.

- **Pour le praticien**, c'est ce qui permet de tester des scénarios « et si » sur la configuration exacte du jour, et non plus dans un environnement de formation figé.
- **L'irritant traité** est celui de la décision opérationnelle prise sur la seule expérience individuelle, sans possibilité de simuler les conséquences d'un choix avant de l'engager.
- **Le risque associé** est la *fausse assurance* : un jumeau n'est jamais plus fiable que le modèle physique qui le sous-tend, et les divergences modèle/réel s'accumulent silencieusement au fil du temps si le calage n'est pas régulièrement repris.

2.1.5 IA symbolique et bayésienne

Famille qui repose sur des règles explicites, des graphes de connaissances ou des réseaux de probabilités conditionnelles — par opposition aux familles purement apprenantes.

- **Pour le praticien**, c'est ce qui rend formalisables les arbres d'événements, les arbres de défaillance et les chaînes causales que la communauté de la sûreté manipule depuis quarante ans, en les rendant interrogeables et actualisables dynamiquement.
- **L'irritant traité** est celui de l'étude de dangers qui vit dans un classeur et n'est plus relue entre deux révisions décennales.
- **Le risque associé** est la *rigidité* : le modèle ne sait que ce que l'expert lui a dit de savoir, et le coût de mise à jour peut décourager l'actualisation.

2.1.6 Réalité virtuelle et augmentée pilotée par IA

Famille qui combine des environnements immersifs avec une adaptation algorithmique au profil et aux actions de l'utilisateur.

- **Pour le praticien**, c'est ce qui élargit la formation à des situations que le simulateur pleine échelle ne sait pas reproduire : coactivité complexe, environnements dégradés étendus, scénarios multi-acteurs avec conflit de priorités.
- **L'irritant traité** est celui de la formation à la situation rare — celle qu'on ne rencontre jamais en service, mais qu'il faut savoir gérer le jour où elle survient.
- **Le risque associé** est l'*illusion de maîtrise* : l'environnement virtuel développe les compétences individuelles sans remédier aux fragilités des compétences collectives, celles qui se construisent dans la coordination et le débat partagés.

Point d'attention : Une huitième famille — la robotique autonome ou semi-autonome — soustrait physiquement l'humain au risque (drones d'inspection, robots en zone irradiée, exosquelettes, télé-opération). Elle ne figure pas dans le tableau qui suit parce qu'elle agit matériellement à la place de l'intervenant et non sur l'information qu'il exploite.

2.2 Cinq capacités opérationnelles documentées

Les sept familles décrites ci-dessus n'agissent pas isolément. Elles se combinent en cinq capacités opérationnelles qui structurent l'offre d'IA aujourd'hui documentée en industrie à risque — et c'est sous l'angle de ces capacités, plutôt que des familles, que le praticien rencontre concrètement l'IA. Derrière le terme générique « intelligence artificielle » se cachent des familles technologiques très différentes, dont les niveaux de maturité vont du déploiement industriel large à l'expérimentation. Pour le praticien, ce qui compte n'est pas la technique sous-jacente, mais la capacité opérationnelle qu'elle rend possible.

Point d'attention : Une précision méthodologique est nécessaire. Dans la littérature, l'efficacité de l'IA est le plus souvent mesurée par la performance du modèle — précision, rappel, taux de détection, réduction des faux positifs. Ces indicateurs ne doivent pas être confondus avec une performance de sécurité. Détecter mieux ne signifie pas nécessairement prévenir mieux. Entre le signal produit par l'IA et la réduction effective du risque, il existe une chaîne sociotechnique : interprétation, priorisation, décision, action, vérification, apprentissage. C'est cette chaîne, plus que le modèle seul, qui détermine l'effet réel sur la sécurité.

2.2.1 Détecter des anomalies dans des données de capteurs

Cette capacité repose principalement sur le **Machine Learning** supervisé et non supervisé, et de plus en plus sur le **Deep Learning** pour les signaux complexes. Des algorithmes d'apprentissage automatique analysent les séries temporelles issues des capteurs de vibration, de température, de pression ou de courant pour repérer les signes précoces de dégradation. C'est la maintenance prédictive — l'usage le plus mature du **Machine Learning** en industrie, déployé à grande échelle depuis plus de dix ans.

2.2.2 Extraire du sens dans des textes

Cette capacité mobilise le **NLP classique** (text mining) et, depuis trois ans, les **grands modèles de langage (LLM)**. Le traitement automatique du langage naturel permet de classer, synthétiser et interroger des corpus textuels volumineux — rapports d'incidents, analyses de retour d'expérience, procédures, rapports d'inspection. Le NLP permet désormais de classer automatiquement des corpus textuels d'événements à grande échelle, avec des performances qui restent dépendantes du contexte sectoriel et de la qualité des annotations d'entraînement.

C'est la seule vraie discontinuité technologique récente : les **LLM génératifs** changent la nature de l'interaction entre l'humain et la donnée en permettant l'interrogation en langage naturel. Mais le risque d'hallucination — des réponses plausibles, mais fausses — impose une validation humaine systématique.

2.2.3 Reconnaître des situations visuelles

Cette capacité repose sur la **vision par ordinateur**, fondée sur du **Deep Learning** convolutif. Elle produit de l'information structurée à partir de flux vidéo : détection du port des équipements de protection individuelle, franchissement de zones interdites, fuites visibles, écarts de conformité gestuelle. Cette information peut alimenter un processus de décision humain (tableau de bord, remontée asynchrone, alimentation d'un REX) — usage qui relève du périmètre de ce *Cahier*, y compris lorsque sa finalité est la protection de l'homme sur un chantier.

Dans son usage informationnel, la **vision par ordinateur** améliore la détection, mais la surveillance permanente peut dégrader la confiance et la culture de signalement si elle est perçue comme un outil de contrôle plutôt que comme un outil d'apprentissage collectif.

2.2.4 Simuler des scénarios

Cette capacité combine les **jumeaux numériques** et la **réalité virtuelle pilotée par IA**, avec un héritage long de la simulation classique. Les industries à haut risque ont une longue familiarité avec la simulation : les simulateurs pleine échelle qualifient depuis quarante ans les équipes de conduite sur les transitoires et les scénarios accidentels **[Bainbridge 1983]**. Ce que l'IA ajoute à cet héritage est d'une autre nature.

Les **jumeaux numériques**, en s'alimentant en continu des données capteurs de l'installation réelle, permettent de tester des scénarios « et si » en ligne et en contexte — non plus dans un environnement de formation fermé, mais sur la configuration exacte du jour.

La **réalité virtuelle pilotée par IA** élargit la formation à des situations que le simulateur pleine échelle ne peut reproduire : coactivité complexe, environnements dégradés étendus, scénarios multi-acteurs. Ces deux technologies restent encore peu matures en contexte de sécurité industrielle.

Un jumeau numérique n'est pas plus fiable que le modèle physique qui le sous-tend, et les divergences modèle/réel s'accumulent silencieusement au fil du temps [IAEA TECDOC-2031, 2023]. La réalité virtuelle développe les compétences individuelles sans remédier aux fragilités des compétences collectives — celles qui tiennent dans la coordination, le débat et la décision partagée.

2.2.5 Soustraire l'humain aux tâches dangereuses

La robotique autonome ou semi-autonome — drones d'inspection, robots en espace confiné, robots en zone irradiée — supprime directement l'exposition humaine au risque. L'apport en sécurité est ici le plus immédiat et le moins contestable. Mais le déploiement complètement autonome reste prématuré dans les environnements non structurés ; l'approche semi-autonome avec opérateur qualifié demeure la recommandation dominante.

Ce qui se déploie massivement aujourd'hui relève majoritairement d'IA-outil — des solutions packagées, vendables par un fournisseur d'équipement, démontrables par un indicateur de retour sur investissement à court terme : maintenance prédictive, optimisation énergétique, qualité produit, vidéosurveillance d'intrusion. Ce qui peine à se déployer relève majoritairement d'IA-système porteur d'un modèle de sécurité — des constructions sur mesure exigeant l'intégration de données fragmentées issues du REX, des quasi-accidents et de l'observation terrain, dont le modèle de sécurité doit être explicité avant l'entraînement, et dont la valeur ne se mesure ni sur un trimestre ni sur un seul indicateur. Ces systèmes-là sont rarement vendus clés en main ; ils sont à construire.

On peut anticiper que cette asymétrie ne se résorbera pas spontanément par jeu du marché. Elle exige un investissement délibéré du donneur d'ordre — non seulement budgétaire, mais aussi en capacité à spécifier ce qu'on attend d'un système IA en sécurité, à l'évaluer, et à l'intégrer à une architecture sociotechnique qui en porte le modèle de sécurité. C'est l'une des dimensions de la fonction d'**architecte** que ce *Cahier* explore.

2.3 Ce que la combinaison de ces briques rend possible — les finalités pour la sécurité

Prises séparément, les cinq capacités qui précèdent restent des briques techniques. C'est leur combinaison qui peut produire des effets utiles à la maîtrise des risques — à condition que l'architecture le permette et que les limites développées dans la section suivante soient traitées. On peut anticiper que les finalités informationnelles suivantes structureront les usages de l'IA en industrie à risque :

- **Anticiper les risques** : La combinaison de la détection d'anomalies sur données capteurs (**ML/DL**), de l'analyse textuelle des rapports d'intervention antérieurs (**NLP / LLM**) et de la vision par ordinateur appliquée à l'état des installations ouvre la possibilité d'identifier plus tôt des situations à risque — que ce soit la dérive d'un équipement, la récurrence d'un type d'écart ou la convergence de signaux faibles jusque-là dispersés.
- **Éclairer la décision humaine** : Croisée avec le contexte d'une activité en cours, l'information produite par ces briques peut enrichir la représentation qu'un opérateur ou un préparateur se fait de la situation — faire apparaître un précédent pertinent, un écart par rapport à un référentiel, une hypothèse à vérifier. Cette finalité suppose une architecture humain-IA qui désynchronise le tempo algorithmique et le tempo de décision.
- **Protéger l'homme par la supervision informationnelle** : La **vision par ordinateur** et l'analyse de flux continus peuvent produire de l'information utile à la protection des intervenants sur des chantiers à risque — détection du port d'EPI adapté au risque, franchissement de zones interdites, présence à proximité d'un équipement en mouvement. Cette finalité est développée au chapitre 4 (Q4) avec la tension qu'elle porte entre protection et surveillance. Elle se distingue de la finalité d'éloignement physique par la robotique, qui relève d'un autre périmètre.
- **Enrichir le retour d'expérience pour améliorer la préparation** : L'extraction de sens dans les corpus textuels d'événements (**NLP / LLM**), l'analyse de la récurrence des écarts et le croisement avec les données d'exploitation peuvent transformer le REX — de compte-rendu d'événements isolés en capacité d'apprentissage à l'échelle d'un parc ou d'une filière. Cette finalité est celle dont les bénéfices sont les plus différés, mais aussi, peut-être, les plus structurants pour la culture de sécurité.

Point clé

Ces finalités précédentes ne sont pas indépendantes — chacune peut nourrir les autres, et c'est précisément cette circulation qui fait la valeur d'un déploiement intégré. Elles structureront les six questions opérationnelles du chapitre 4.

▷ **Éloigner l'homme du risque par la robotique autonome ou semi-autonome** : Cette cinquième finalité relève d'un autre périmètre, celui des IA qui agissent matériellement à la place de l'intervenant. Son apport en sécurité est direct, mais suppose un traitement technique distinct ; ce point ne sera pas traité dans ce document.

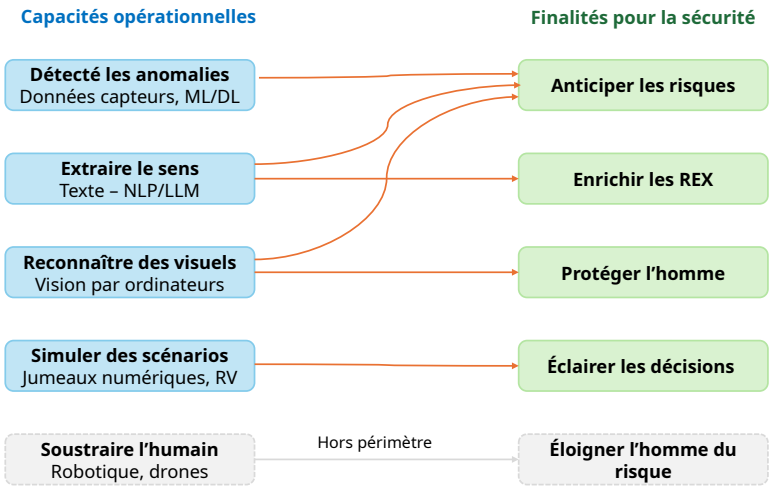

Fig. 2.2 Des capacités aux finalités pour la sécurité industrielle.

2.4 Ce que l'IA ne sait pas faire

La cartographie des cinq capacités et des finalités pourrait laisser entendre que l'IA constitue une extension homogène et linéaire de la capacité d'action de l'organisation. Cinq incapacités structurelles imposent de tracer une frontière nette.

Le progrès technologique en repoussera certaines et en atténuera d'autres ; ce n'est pas le pari sur lequel repose ce chapitre. Ce qui demeure, quel que soit le niveau de performance atteint, c'est qu'un acte reste à chaque fois non transférable à la machine : instituer une décision, surveiller la pertinence d'un outil, faire face à l'inédit, établir le lien causal qui désigne les barrières, assumer un arbitrage. L'enjeu n'est pas ce que l'IA saura faire demain, mais ce dont elle ne peut pas être tenue responsable — et que l'organisation doit, par conséquent, continuer de porter.

Une précision conditionne la lecture de tout le chapitre. Aucune de ces frontières n'interdit le déploiement de l'IA, et aucune n'appelle la même réponse. Il faut distinguer ce qui pourrait être surmonté techniquement de ce qui ne peut être que contourné organisationnellement. Une IA ne portera pas la responsabilité d'un arbitrage ; une organisation peut en revanche l'organiser — l'attribuer, la tracer, la rendre effective[2]. C'est précisément l'objet de ce *Cahier* : non pas attendre que l'IA franchisse ces frontières, mais construire l'architecture humain-IA qui les tient.

Il importe enfin de distinguer ces incapacités, propriétés structurelles de la relation humain-IA, des risques d'usage qui résultent de la manière dont l'organisation déploie l'outil. Les seconds seront traités au fil des chapitres suivants. Les premières définissent la frontière au-delà de laquelle l'humain reste seul.

[2] Consulter l'article 26 du règlement européen sur l'IA et le rapport du Contrôleur européen de la protection des données **[EDPS 2025]**.

L'IA ne sait pas instituer ce qu'elle propose

Une recommandation algorithmique ne porte pas, en elle-même, d'autorité organisationnelle. Dans le périmètre informationnel qui est celui de ce Cahier, l'IA produit une information qu'un humain interprète, valide et institue : l'acte de décision reste, par construction du périmètre, un acte humain. Des dispositifs où la machine institue effectivement la décision existent — pilotes automatiques, systèmes d'arrêt d'urgence, freinage automatique, robotique autonome — mais relèvent d'un autre périmètre.

Lorsque cette frontière n'est pas tenue, le biais d'automation s'installe : les décideurs accordent un poids excessif aux recommandations algorithmiques, au point de leur conférer une légitimité que ni leur source ni leur méthode de production ne justifient. La signature humaine devient alors un acte formel de ratification d'une décision qui n'a pas été instituée, ce qui constitue un **blanchiment de légitimité** [Davies 2025] (*cf.* la figure 2.3). L'enjeu n'est pas de doter l'IA d'une autorité qu'elle n'a pas, mais de préserver l'acte d'institution comme un acte humain pleinement assumé. Cela suppose de distinguer trois moments — la recommandation algorithmique, l'examen humain, l'institution organisationnelle — et de refuser leur fusion en un geste unique.

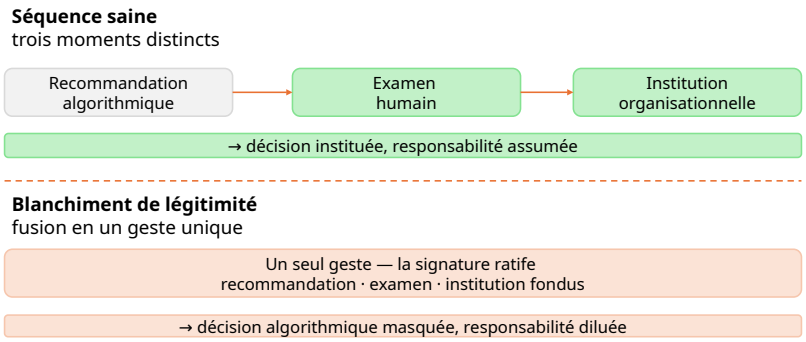

Fig. 2.3 Blanchiment de légitimité — séquence saine et effondrement.

L'IA ne sait pas à elle seule reconnaître quand elle ne sait plus

Un modèle apprenant performant sur la distribution de données pour laquelle il a été entraîné peut s'effondrer silencieusement quand cette distribution dérive — sans le signaler par lui-même. Le contexte change, les données d'entrée deviennent obsolètes, les hypothèses de conception ne tiennent plus, et le modèle continue de produire des résultats cohérents en apparence, parce qu'il n'a pas, en propre, de méta-conscience de ses propres limites de validité.

Des techniques de détection de hors-distribution se sont développées depuis 2017 et permettent, dans certaines conditions, de signaler qu'une donnée d'entrée s'écarte du périmètre d'entraînement. Ces techniques restent toutefois imparfaites et elles ne fonctionnent que lorsqu'un superviseur humain a défini en amont ce qui constitue une dérive significative. La machine ne sait pas, à elle seule, qu'elle ne sait plus.

La distinction est nette avec la panne technique. La panne est identifiable — arrêt, erreur, indisponibilité — et enclenche les procédures de bascule. La perte d'utilité opérationnelle est d'une autre nature : le système fonctionne, produit des résultats apparemment cohérents, mais n'apporte plus d'aide pertinente à l'activité réelle. L'organisation continue de s'appuyer sur un outil devenu inadapté parce qu'il a toujours bien fonctionné.

Pour le praticien, la conséquence est nette : un système d'IA qui fonctionne techniquement n'est pas un système qui protège. La surveillance de la pertinence opérationnelle ne peut pas être déléguée au modèle lui-même. Elle suppose un regard humain entretenu, appuyé sur une compréhension du métier que la routine d'usage tend précisément à éroder.

L'IA ne sait pas opérer dans l'inconnu

L'incapacité précédente porte sur la méta-connaissance — l'IA ne sait pas qu'elle a quitté son domaine. Celle-ci porte sur la performance — même informée qu'elle l'a quitté, elle ne saurait pas y opérer. Un **modèle apprenant** généralise mal hors de la distribution de données sur laquelle il a été entraîné. Face à une situation jamais rencontrée, il produit des sorties plausibles, mais arbitraires. C'est un argument central des travaux récents sur la généralisation hors distribution [Bengio et al. 2021 ; Mitchell 2019]. Un **modèle apprenant** n'invente pas — il interpole et, dans une certaine mesure, extrapole à partir de ce qu'il a vu. L'inconnu, par définition, n'a pas été vu.

Or les accidents majeurs des industries à haut risque — Three Mile Island, Bhopal, Texas City — sont par définition hors de la distribution d'entraînement de tout modèle. C'est précisément leur caractère inédit qui les rend graves : ils violent les hypothèses sous-jacentes au système de barrières. Un **modèle entraîné** sur l'historique des incidents et quasi-incidents d'une installation peut détecter les variantes des situations connues. Il ne peut pas anticiper la situation que personne n'a anticipée.

C'est exactement la zone où l'opérateur expérimenté reste irremplaçable. Sa capacité à reconnaître qu'on est dans un territoire jamais cartographié, à mobiliser une analogie opérationnelle nourrie par des années d'expérience incarnée, à dialoguer avec d'autres opérateurs pour construire un sens partagé d'une situation inédite n'est pas réplicable par un modèle apprenant a à dialoguer avec d'autres opérateurs pour construire un sens partagé d'une situation inédite n'est pas réplicable par un modèle dont la compétence dérive de sa distribution d'entraînement.

L'IA ne sait pas, seule, établir un lien de causalité

Un modèle apprenant entraîné sur des données observationnelles apprend des corrélations, pas des relations causales. Il peut prédire qu'un événement X est statistiquement associé à un événement Y, mais il ne peut pas répondre, à partir de ces seules données, à la question « qu'est-ce qui se passerait si on intervenait sur X ? ». Cette distinction n'est pas un défaut de maturité que de meilleures données combleraient : elle est épistémologique, identifiée par [Pearl et Mackenzie 2018] et résumée par sa hiérarchie des trois niveaux — observation (ce que ces modèles font bien), intervention et contrefactuel (ce qu'ils ne font pas à partir de l'observation seule). Des méthodes causales existent et peuvent, lorsqu'on leur fournit un modèle causal explicite ou des données interventionnelles, monter ces étages. Mais elles déplacent l'acte, elles ne le suppriment pas : quelqu'un doit poser le modèle causal, et surtout décider quelles interventions comptent pour les barrières. C'est cet acte-là qui ne se délègue pas.

Cette limite vaut aussi pour les grands modèles de langage récents, dont les capacités de raisonnement apparent peuvent donner le sentiment qu'ils maîtrisent la causalité. Ils en récitent fluidement le vocabulaire — « X cause Y », « si X n'avait pas eu lieu, alors Y… » — parce que ces énoncés sont massivement présents dans leurs corpus d'entraînement. Mais les évaluations récentes [Zečević et al. 2023 ; Jin 2025] convergent : sur des situations nouvelles, hors corpus d'entraînement, leur performance s'effondre. Ils restent à ce que Pearl appelle l'étage 1 de l'échelle de causalité — l'observation.

Le risque pour le praticien n'a pas diminué, il s'est déplacé : la fluidité apparente de la réponse rend désormais plus difficile la reconnaissance du fait qu'aucun raisonnement causal n'a eu lieu. Pour la sécurité industrielle, l'enjeu est massif. Le REX ne sert pas à prédire les prochains accidents — il sert à comprendre quelles barrières installer pour les empêcher. Ce qui compte est l'intervention, pas la prédiction. Or l'intervention suppose un raisonnement causal que ces modèles ne portent pas.

L'IA ne sait pas assumer un compromis entre objectifs contradictoires

Toute décision de sécurité industrielle est un arbitrage : sécurité et production, court terme et long terme, individuel et collectif, coût et bénéfice. L'IA peut optimiser une fonction objectif définie à l'avance, mais elle ne peut pas *assumer* un compromis — elle ne peut pas dire *« j'ai privilégié X au détriment de Y, et j'en porte la responsabilité dans tel horizon de finalité »*.

L'optimisation algorithmique et l'arbitrage humain sont deux opérations de nature différente. L'optimisation suppose une fonction objectif stable et mesurable. L'arbitrage suppose une délibération sur ce qui compte, une hiérarchisation entre des valeurs incommensurables, et un acte qui engage celui qui décide. Sans cet acte, il n'y a pas de décision — il y a un calcul.

Ces incapacités tracent ensemble la frontière au-delà de laquelle l'humain reste seul. Elles ne sont pas des arguments contre l'IA — elles sont les conditions à intégrer dans toute architecture humain-IA qui prétend traiter de la sécurité. Une organisation qui déploie l'IA sans avoir nommé ces frontières ne déploie pas un outil ; elle abandonne un territoire de décision sans en avoir conscience. À l'inverse, une organisation qui les nomme et qui les outille — en préservant l'acte d'institution, la surveillance de la pertinence opérationnelle, la préparation aux situations inédites, le choix humain des barrières et l'arbitrage assumé — peut tirer parti des capacités décrites en ouverture du chapitre sans en payer le prix structurel. C'est, au fond, la posture HRO appliquée à l'IA : ni défiance, ni délégation, mais une vigilance qui sait ce qu'elle ne confie pas.

Note méthodologique : Ces cinq incapacités ne sont pas exhaustives. Ce sont celles qui paraissent les plus structurantes pour la sécurité industrielle ; un Cahier de plus large portée pourrait en mentionner d'autres — la difficulté à dialoguer avec une situation au sens de **[Schön 1983]**, l'absence d'intentionnalité au sens de **[Smith 2019]**, la difficulté à exercer un jugement éthique. Le choix opéré ici est celui de la pertinence directe pour le praticien.

2.5 À quelles conditions un déploiement est-il soutenable ?

Les cinq limites qui précèdent ne sont pas des raisons de renoncer à l'IA. Elles sont des conditions à intégrer dans tout projet de déploiement. En pratique, cinq prérequis se dégagent de l'expérience des secteurs les plus avancés :

- **La qualité des données prime sur la sophistication de l'algorithme**. Un modèle d'IA ne vaut que ce que valent ses données d'entrée. Les fragilités 4, 5 et 6 du chapitre précédent ont montré que les données de sécurité sont fragmentées, hétérogènes et souvent de qualité inégale. Déployer de l'IA sur des données médiocres ne produit pas de l'intelligence ; cela automatise les biais existants. Tout projet doit commencer par un travail de nettoyage, de structuration et de gouvernance des données — un travail peu glamour, mais déterminant.
- **L'intégration aux systèmes existants est un défi d'ingénierie, pas un problème d'algorithme**. Les installations à risque fonctionnent avec des équipements hétérogènes, de différents âges, avec des systèmes d'information qui n'ont pas toujours été conçus pour communiquer entre eux. Extraire les données, les normaliser, assurer une connectivité fiable et sécurisée suppose parfois de moderniser une partie de l'infrastructure — un investissement non négligeable que la direction devra arbitrer en fonction des gains attendus.
- **L'acceptabilité par les collectifs de travail conditionne l'efficacité**. Un système d'IA rejeté par les opérateurs ou perçu comme un outil de surveillance ne produit pas d'amélioration durable. L'intégration réussie suppose une conception centrée sur l'humain, une transparence sur ce que **le système** observe et décide, et une implication des utilisateurs dès la conception. La question éthique — jusqu'où est-il acceptable de traquer les comportements d'un salarié par caméra, même pour sa sécurité ? — doit être traitée comme un sujet de gouvernance, pas comme un détail de déploiement.
- **Le cycle de vie d'un modèle apprenant est continu, pas ponctuel**. Contrairement à un logiciel classique, dont le comportement est inscrit dans le code et ne change qu'au prix d'une modification explicite de ce code, un modèle apprenant tire son comportement des données sur lesquelles il a été entraîné. Sa performance se dégrade silencieusement à mesure que les données opérationnelles s'écartent de celles d'entraînement. Le maintenir en condition opérationnelle suppose un cycle régulier de réétalonnage, de revalidation et

de redéploiement, dans lequel chaque ré-entraînement peut produire des effets non locaux et partiellement imprévisibles.

L'organisation est confrontée à une injonction paradoxale : devenir plus agile pour suivre cette dynamique, tout en conservant la rigueur démonstrative que la sûreté exige. C'est un paradoxe que les industries à haut risque n'ont pas encore résolu.

- ▷ **La cybersécurité est une condition non négociable**. Chaque connexion créée pour alimenter un modèle d'IA élargit la surface d'attaque. La fragilité 3 du chapitre précédent a montré que la numérisation de l'entreprise étendue crée des vulnérabilités nouvelles. La directive NIS2 étend les obligations de cybersécurité aux sous-traitants critiques. Un déploiement d'IA qui ne s'inscrit pas dans une architecture de sécurité informatique robuste est un risque, pas un progrès.

La confidentialité et la souveraineté des données conditionnent les usages autorisés

Confier des données d'exploitation, d'analyse d'événements ou de REX à un système d'IA soulève trois questions qui se confondent souvent dans le débat public, mais qu'il faut distinguer pour agir.

- ▷ **La première est celle du canal d'échange**. Saisir un contenu sensible dans l'interface grand public d'un LLM — ChatGPT, Claude, Gemini — revient à le transmettre au fournisseur, avec un régime de conservation, d'accès et d'usage qui dépend de l'offre souscrite : compte individuel gratuit, API avec clause *no-training*, ou offre entreprise sur environnement dédié. La majorité des fuites documentées à ce jour tiennent moins d'une propriété technique des **LLM** que d'une absence de doctrine d'usage dans les organisations.
- ▷ **La deuxième concerne le modèle entraîné sur des données internes.** Une croyance répandue veut qu'un modèle *fine-tuné* « recracherait » ses données d'entraînement à qui saurait poser la bonne question. La réalité est plus nuancée : des attaques d'extraction existent **[Carlini et al. 2021 ; Nasr et al. 2023]**, mais elles exigent un accès spécifique au modèle et produisent des résultats partiels. On n'accède pas à une base de REX confidentielle en interrogeant un LLM comme un moteur de recherche. La vigilance est justifiée ; la confusion avec un risque de fuite triviale ne l'est pas.
- ▷ **La troisième relève de la souveraineté.** Les modèles les plus performants sont aujourd'hui conçus et opérés par des acteurs américains ou chinois ; le CLOUD Act [2018] permet aux autorités fédérales américaines d'exiger la communication de données relevant de leur juridiction, y compris lorsque ces données sont stockées hors des États-Unis. Pour une installation nucléaire, un site Seveso ou un opérateur d'importance vitale, cette exposition n'est pas théorique.

Les alternatives — modèles européens sur infrastructure souveraine, modèles hébergés en interne — existent, mais présentent à ce jour un déficit de performance sur les tâches analytiques les plus exigeantes. L'arbitrage entre performance et souveraineté n'est pas neutre : il relève du comité de direction, non de la seule fonction technique.

La certification de ce qui peut s'apparenter à une « boîte noire »

Une dernière condition, qui dépasse le périmètre technique de ce chapitre, concerne la certification réglementaire (ou démonstration de sûreté) des fonctions de sûreté/sécurité confiées à l'IA. Comment un exploitant démontre-t-il à son autorité de contrôle qu'un système apprenant est fiable, dans un cadre de gouvernance conçu pour des systèmes déterministes ? Cette question sera traitée dans le chapitre suivant, à travers la grille de décision qui distingue ce que l'on peut déléguer, ce que l'on doit assister et ce qu'il faut sanctuariser (*cf.* § 3.3).

2.6 Ce que l'IA change dans la manière de penser la sécurité

Au-delà des capacités et des limites, l'IA introduit un changement plus profond : elle oblige à redéfinir la distribution des rôles entre l'humain et la machine dans la maîtrise des risques.

Les approches classiques de la sécurité reposent sur une distinction implicite : les machines exécutent, les humains décident. L'IA brouille cette frontière. Un algorithme de maintenance prédictive ne se contente pas d'exécuter un calcul ; il formule une recommandation qui influence directement une décision opérationnelle — reporter ou avancer une intervention, modifier un planning, réévaluer un risque. L'opérateur reste formellement décideur, mais le cadre dans lequel il décide est désormais structuré par un système qu'il ne comprend pas toujours et qu'il ne peut pas auditer lui-même.

Cette redistribution pose une question centrale pour la suite de ce document : pour chacune des activités de maîtrise des risques — préparation, détection, analyse, pilotage — qu'est-ce que l'on peut confier à l'IA, qu'est-ce que l'on doit garder sous contrôle humain, et qu'est-ce qu'il faut protéger de toute interférence algorithmique ? Cette grille de décision — *déléguer, assister, sanctuariser* — sera développée dans le chapitre suivant.

Une chose peut déjà être dite. L'IA informationnelle ne modifie pas les fondamentaux physiques et anthropologiques de la sécurité industrielle — les accidents obéissent toujours aux mêmes lois, les opérateurs restent des humains avec leurs limites cognitives, les organisations restent traversées par les mêmes jeux d'acteurs. Mais l'IA modifie la manière dont ces fondamentaux s'expriment, et parfois profondément. Ce que la littérature en facteurs humains documente depuis **[Bainbridge 1983]**, et que les travaux récents sur le biais d'automation actualisent dans le contexte de l'IA **[Parasuraman et Manzey 2010 ; Strauch 2018]**, peut se résumer en trois déplacements pour le praticien :

- **Elle déplace le rapport de l'opérateur au doute, à la vigilance et au jugement** — non par manque de formation, mais parce que le mécanisme est attentionnel : la vigilance s'érode silencieusement à mesure que le système fonctionne bien, et la compétence du jugement se dégrade faute d'être pratiquée **[Bainbridge 1983 ; Parasuraman et Manzey 2010]**.
- **Elle réorganise les zones d'autorité épistémique entre ceux qui maîtrisent l'outil et ceux qui n'y ont pas accès**, en déplaçant le centre de gravité de l'expertise — d'une expertise du métier vers une expertise de l'interface entre métier et système.
- **Elle installe une asymétrie cognitive entre les entreprises, les métiers et les générations** — entre celles qui ont les moyens d'industrialiser ces dispositifs et celles qui les subissent, entre les métiers où l'IA mature et ceux où elle reste expérimentale, entre les générations formées avec ces outils et celles qui les ont rencontrés tard.

Ce qu'elle change surtout, **c'est la capacité de l'organisation à traiter de l'information à une échelle et à une vitesse que l'humain seul ne peut pas atteindre**. C'est précisément là que se situe la réponse aux dimensions informationnelles des six fragilités — si et seulement si le système sous-jacent est suffisamment sain pour en tirer profit.

Question ouverte

Le déploiement de l'IA industrielle va-t-il progressivement transférer aux machines une partie du pilotage de la sécurité aujourd'hui exercé par les humains ? Et si oui, les modèles de FOH qui fondent le management de la sécurité depuis trente ans sont-ils encore adaptés ?

2.7 De la performance technique à la transformation effective du risque : l'enjeu de la chaîne d'action

Les revues récentes sur l'IA appliquée à la sécurité industrielle conduisent à une conclusion prudente. Les capacités techniques de l'IA sont désormais bien documentées : détection d'anomalies, reconnaissance visuelle, analyse de textes, prédiction de situations dangereuses, classification d'événements, assistance à l'évaluation dynamique des risques. Mais la preuve de leur effet direct sur la réduction des accidents, maladies professionnelles ou événements majeurs reste encore limitée.

Point clé

La littérature mesure beaucoup la performance des modèles ; elle mesure beaucoup moins la transformation effective du risque.

Cette distinction est décisive pour le praticien. Une IA qui détecte mieux ne prévient pas nécessairement mieux. Entre le signal produit et la réduction effective du risque, il faut une chaîne complète : qualité des données, interprétation métier, priorisation, décision, action terrain, vérification d'efficacité et apprentissage organisationnel. C'est cette chaîne qui transforme une capacité algorithmique en performance de sécurité. Sans elle, l'IA peut produire davantage de signaux, davantage de tableaux de bord et davantage d'alertes, sans rendre l'organisation plus sûre.

3

Quel modèle de sécurité apprenez-vous à l'IA ?

Le chapitre précédent a montré ce que les systèmes IA savent faire et ce qu'ils ne savent pas faire. Mais ces capacités techniques ne disent rien de la question centrale : **sur quoi ces systèmes vont-ils travailler ?** Un algorithme de maintenance prédictive peut être techniquement excellent et dégrader la sécurité s'il a été entraîné sur des données qui reflètent un modèle de sécurité défaillant. Un système de détection de signaux faibles peut être performant et produire de la complaisance s'il détourne l'attention de ce qu'il ne voit pas. La technologie ne décide pas de la sécurité ; c'est le modèle de sécurité que portent — implicitement ou explicitement — les systèmes IA déployés qui la décide.

Ce chapitre développe la thèse centrale du document : avant de déployer un système IA dans un dispositif de sécurité, on doit expliciter le modèle de sécurité sur lequel il va s'appuyer — et décider ce qu'on lui confie, ce qu'on garde sous contrôle humain et ce qu'on protège de toute interférence algorithmique. Cette exigence concerne, utilisant les termes introduits au chapitre 2, les systèmes d'IA-système porteur d'un modèle de sécurité — c'est-à-dire ceux qui interviennent, ou prétendent intervenir, dans la résolution des fragilités identifiées au chapitre 1.

Qu'on en ait ou non conscience, la sécurité repose sur un modèle, c'est-à-dire une référence partagée par tous du niveau de risque acceptable et de ce qui doit être mis en œuvre dans le but d'éviter les accidents. Le modèle de sécurité est la réponse collective à une question exigeante : ***qu'est-ce qui fera qu'il n'y aura pas d'accident grave ou majeur dans notre organisation ?***

Ces pratiques, outils et comportements (qui peuvent être observés) sont basés sur des représentations mentales, des croyances et des valeurs (qui ne sont pas directement observables).

Définition

Un modèle de sécurité

Un modèle de sécurité, tel que nous utilisons ce terme dans ce document, est un référentiel opérationnel explicite et partagé qui relie les situations à haut potentiel de gravité (SHPG) à un système de défense en profondeur, en combinant des barrières de nature technique, système de management (SMS) et FOH. Il précise les pilotages locaux et globaux, les précurseurs/perturbateurs et leurs parades. Surtout, il formalise les équilibres structurants du dispositif (comme les arbitrages assumés par l'organisation).

3.1 Les trois modèles de sécurité — et celui que l'IA apprend

Toute organisation à haut risque fonctionne avec trois modèles de sécurité, qu'elle le sache ou non.

3.1.1 Le modèle de sécurité prescrit (voulu)

Le **modèle prescrit** est ce qui est écrit : les procédures, les analyses de risques, les démonstrations de sûreté, les systèmes de management, les règles qui sauvent. Il résulte d'un travail d'élaboration et d'arbitrage par lequel l'organisation traduit, hiérarchise et met en cohérence un cadre normatif hétérogène — réglementation, normes, exigences clients, valeurs internes — dans un système qui lui est propre. C'est ce modèle que l'organisation montre à son autorité de contrôle et qui peut faire l'objet d'une certification. Il exprime, qu'on le veuille ou non, la politique de sécurité de l'organisation.

Encore faut-il qu'il soit lisible. Dans une organisation mature, le modèle prescrit est explicite : on peut nommer les principes qui le structurent, identifier les arbitrages qui ont été faits, retrouver dans les documents la trace des choix de l'organisation.

Mais cette situation est loin d'être la règle. Souvent, le modèle prescrit existe sans avoir été véritablement élaboré comme tel : il est dilué dans la masse documentaire du système de management, dispersé entre des centaines de procédures, des analyses de risques cumulées, des consignes superposées au fil des années. Le modèle est *là*, mais personne ne saurait le formuler. Il est implicite, parfois même invisible à ceux-là mêmes qui le portent. Cette opacité n'est pas neutre : elle empêche le débat sur les choix de sécurité, elle fragilise la transmission entre générations de managers, et elle rend tout dialogue avec la ligne managériale et le terrain plus difficile — car comment confronter un modèle porté à un modèle prescrit qui n'est lui-même pas explicité ?

3.1.2 Le modèle managérial porté

Le **modèle managérial porté** est ce que la ligne managériale a construit, dans la durée, comme réponse à la question : *qu'est-ce qui fait qu'il n'y aura pas d'accident grave et mortel ?* Cette construction collective s'élabore par croisement des croyances héritées (formations, trajectoires, cultures professionnelles), du prescrit, et de l'expérience partagée du collectif (événements, REX, succès, échecs). Elle est faite de représentations qui guident l'action, les arbitrages quotidiens, ce qui est valorisé ou ignoré, ce qui est regardé ou laissé dans l'ombre. Elle inclut aussi des croyances qui peuvent fragiliser la sécurité : *« ça n'arrivera pas chez nous »*, *« on a toujours fait comme ça »*, *« le sous-traitant connaît son métier »*.

Ce modèle managérial porté n'est pas toujours unifié. Dans un collectif managérial stable, qui travaille ensemble depuis longtemps, il est en général homogène[1] : interrogés séparément sur ce qui empêche l'accident, les managers convergent. Cette homogénéité est un marqueur de maturité du collectif — qu'elle aille dans le bon sens ou dans le mauvais. À l'inverse, dans une organisation marquée par une forte rotation managériale, par une fusion-acquisition récente, par un fonctionnement multisite peu coordonné, ou par la coexistence de plusieurs cultures professionnelles qui ne se sont pas encore intégrées, le modèle managérial porté peut être disparate.

Plusieurs versions coexistent alors, parfois contradictoires, sans qu'aucune ne domine clairement. Les deux configurations posent des problèmes de nature très différente : un modèle homogène, mais erroné est solide et difficile à faire évoluer ; un modèle disparate ne fournit pas une boussole partagée, et chaque manager pilote la sécurité avec sa propre carte.

[1] Un modèle managérial homogène, mais erroné est solide parce qu'il fonctionne comme une "basic underlying assumption" au sens de **[Schein et Schein 2017]** — une croyance partagée devenue évidente, peu sensible aux faits qui la contredisent. **[Weick 1995]** montre la viscosité propre de ces théories construites par sensemaking collectif. À l'inverse, un modèle disparate signale souvent l'absence des conditions de sécurité psychologique **[Edmondson 1999]** qui permettent à un collectif de confronter ses croyances et de construire une compréhension partagée.

3.1.3 Le modèle de sécurité opérant

Le **modèle opérant** est ce qui se fait effectivement sur le terrain : les adaptations compétentes, les contournements acceptés, les arbitrages implicites entre production et sécurité, les règles appliquées et celles qui ne le sont plus. C'est le travail réel, au sens de l'ergonomie. Il résulte d'une élaboration par l'activité, où les opérateurs croisent l'intention managériale, la matérialité du travail (aléas, variabilité, contraintes, coactivités), leur ingéniosité et les régulations du collectif de travail.

La figure 3.1 représente cette configuration. Les trois modèles devraient théoriquement se superposer parfaitement ; en pratique, le contenu se réduit et se décale à mesure qu'il descend vers le terrain, et des pratiques locales émergent en marge. Cette structure en rétrécissement avec débordement est, dans l'expérience de l'auteur, la configuration la plus fréquente en industrie à haut risque. Ces trois modèles ne coïncident jamais parfaitement. Mais cette divergence n'est pas en elle-même un défaut : elle peut révéler une richesse — un terrain qui a trouvé mieux que le prescrit, des opérateurs qui adaptent intelligemment dans des situations imprévues, une ligne managériale qui priorise ce qui compte vraiment au-delà de ce qui est écrit ; elle peut aussi révéler une vulnérabilité — un prescrit qui ne traduit pas les vraies priorités, une ligne managériale qui ne porte pas ce qui est essentiel, un terrain qui dérive silencieusement.

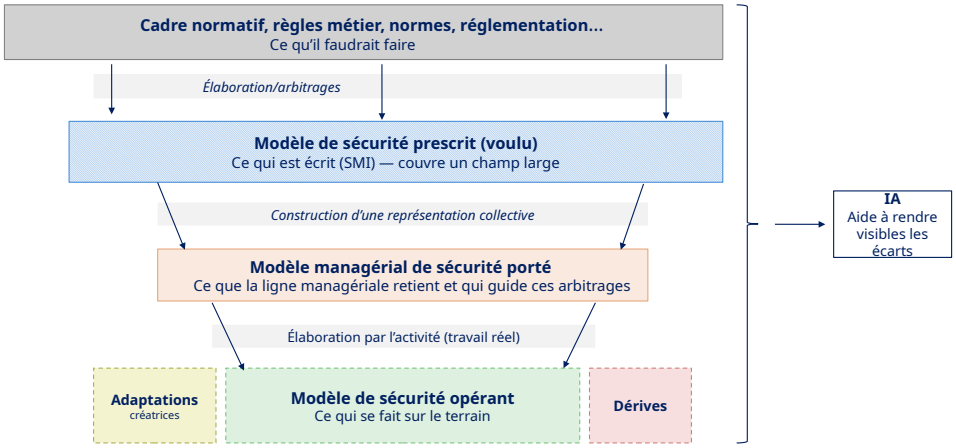

Fig. 3.1 Les trois modèles de sécurité évoqués dans ce chapitre.

Tout l'enjeu de la prévention des accidents graves et mortels est de *qualifier* ces divergences : comprendre pourquoi prescrit, porté et opérant divergent, distinguer les divergences fécondes des divergences qui fragilisent, et agir en conséquence — non pas pour imposer une conformité rigide, mais pour que ce qui est prévu, ce qui est porté et ce qui est fait convergent vers un niveau de sécurité réel et durable.

Cette qualification ne se fait pas au même endroit selon la nature du dispositif. Sur les dispositifs très réglés — consignation, permis de feu, entrée en espace confiné — le modèle de sécurité est largement incrusté dans la règle, et les divergences à qualifier s'observent aux marges. Sur les dispositifs largement gérés — arbitrage d'exploitation en dégradé, gestion d'aléa, retour d'expérience, surveillance par rondes — le modèle de sécurité est explicite et variable, et les divergences peuvent être structurelles. Cette distinction conditionne le poids du travail d'explicitation avant déploiement IA : rapide quand le dispositif est très réglé, déterminant quand il est largement géré.

Point clé

> Tout système IA déployé en sécurité s'appuie, d'une manière ou d'une autre, sur un corpus de données — celles qui ont servi à le construire, celles qu'on lui fournit en contexte, celles qui actualisent ses connaissances. Ce corpus porte un modèle de sécurité, qu'il soit explicité ou non. Si le système s'appuie sur les procédures, il reproduit le prescrit — y compris ses lacunes et ses rigidités. S'il s'appuie sur les données du terrain, il reflète le modèle opérant — y compris ses contournements et ses dérives. Le choix des données sur lesquelles un système IA s'appuie est un choix de modèle de sécurité, même si personne ne le formule ainsi.

Ce point clef prend deux formes très différentes selon la configuration de déploiement.

- Dans le premier cas, l'industriel a la main sur les données : système développé en interne, plateforme spécialisée alimentée par les bases métier, grand modèle de langage utilisé en mode *Retrieval-Augmented Generation* (RAG) sur le corpus documentaire de l'entreprise, fine-tuning ciblé. Le choix des données est alors un choix actif du donneur d'ordre, qui engage explicitement le modèle de sécurité qu'il transmet à la machine.
- Dans le second cas le système est acheté clés en main à un fournisseur : algorithme de maintenance prédictive entraîné sur des données génériques de flotte, dispositif de vision par ordinateur entraîné sur des bibliothèques d'images d'EPI ou de défauts, capteur intelligent embarquant un modèle pré-appris. Le modèle de sécurité — ou plus exactement, le modèle de reconnaissance qui en tient lieu — a été configuré par le fournisseur, sur la base de choix qui ne sont pas ceux du donneur d'ordre, et souvent sans que ces choix aient été formulés explicitement.

Le travail de l'exploitant devient alors différent : il ne s'agit plus de choisir les données, mais de **qualifier ce que le système porte effectivement** — par quels exemples a-t-il été instruit, quelles situations ne sait-il pas reconnaître, quelles définitions implicites de la conformité ou de l'écart hérite-t-il de son entraînement initial. Cette archéologie est plus exigeante que le choix actif des données, parce qu'elle suppose un dialogue avec un fournisseur qui n'a pas toujours intérêt à expliciter ce qu'il a configuré.

3.1.4 Entre explicite et implicite : révéler les véritables modèles de sécurité dans le déploiement de l'IA industrielle

C'est ici que se situe le risque le plus profond — et le plus invisible — du déploiement de l'IA en sécurité industrielle. Si l'organisation n'a pas explicité son modèle de sécurité, si elle ne sait pas quels équilibres structurent ses défenses, si elle ne distingue pas les adaptations créatrices de sécurité des écarts destructeurs, alors elle ne sait pas ce qu'elle enseigne à la machine.

Les valeurs de l'organisation sont encodées dans les données avant le premier algorithme. Les arbitrages implicites — entre production et sécurité, entre conformité et adaptation, entre court terme et long terme — sont déjà dans les données. L'IA ne fera que les révéler ou les amplifier.

3.2 La grille de décision : déléguer, assister, sanctuariser

Une fois le modèle de sécurité explicité, la question suivante est : *que confier à l'IA ?* Olivier Sibony et Éric Hazan, dans *Faut-il encore décider ?* **[Hazan et Sibony 2026]**, proposent un cadre qui, transposé à la sécurité industrielle, structure cette réflexion en trois niveaux de décision.

La grille proposée ici — qu'on appellera grille DAS dans la suite du document — reprend la triade de Sibony et Hazan comme épine dorsale. Elle est complétée d'une **méthode d'arbitrage** structurée autour de cinq critères, et de tests de viabilité qui en vérifient la robustesse dans la durée.

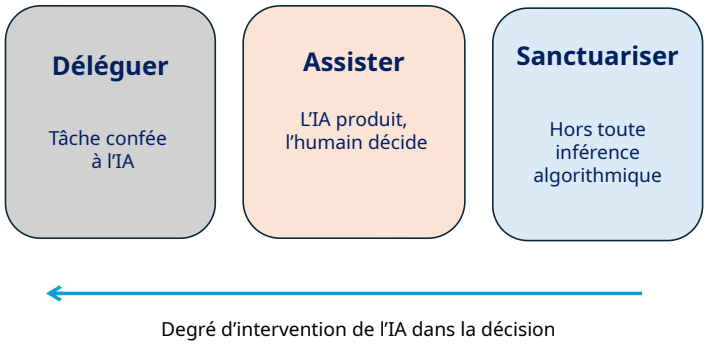

Fig. 3.2 Grille DAS - Déléguer, Assister, Sanctuariser, suivant les travaux de Hazan et Sibony.

3.2.1 La triade : trois positions de décision

La grille DAS distingue trois positions, qui ne sont pas trois façons d'utiliser l'IA, mais trois **façons de structurer une décision** dans laquelle l'IA peut intervenir.

- **Déléguer** : l'organisation accepte que la sortie de l'IA produise directement un effet, sans interposition humaine systématique. Une supervision périodique vérifie que le dispositif reste pertinent, mais chaque sortie individuelle n'est pas validée par un humain. *Exemples : le tri et la classification automatique de signaux capteurs en pré-traitement de maintenance prédictive.*
- **Assister** : l'organisation interpose un humain qui statue, l'IA produisant un intrant à sa décision. C'est le régime pertinent pour les activités où le contexte compte autant que la donnée, où l'ambiguïté est irréductible, et où la décision engage une responsabilité humaine identifiée. La traçabilité y est essentielle : on doit pouvoir distinguer ce que l'IA a détecté, ce que l'humain a prescrit, et ce que le management a décidé. Ce régime recouvre trois niveaux d'assistance, traitées en § 3.2.5. *Exemples : l'aide à l'analyse d'un événement à haut potentiel en suggérant des rapprochements avec des événements antérieurs, l'optimisation des fenêtres de maintenance sur les équipements critiques.*
- **Sanctuariser** : l'organisation interdit à l'IA de produire un intrant à cette décision. Ce qui est sanctuarisé doit être formalisé explicitement, et les compétences nécessaires pour l'assumer doivent être préservées délibérément. Cette préservation a un coût qui doit être anticipé. *Exemples : l'arbitrage entre poursuivre et interrompre une opération en situation dégradée, l'appréciation de la culture juste dans une enquête post-événement.*

Ces trois positions ne s'appliquent pas à « l'IA en général » dans une organisation, mais à chaque fonction identifiée — préparer une intervention, hiérarchiser des signaux, analyser un événement. Une même organisation peut déléguer sur certaines fonctions, assister sur d'autres et sanctuariser ailleurs. C'est précisément l'objet de la grille DAS : aider à choisir, fonction par fonction, où placer le curseur.

Placer le curseur n'est pourtant ni un classement ni un acte définitif. C'est un arbitrage initial, que la grille fixe et rend traçable — mais un régime n'est pas un état stable : c'est une position qui se maintient ou se dégrade, et les tests de viabilité présentés plus loin servent précisément à vérifier qu'elle tient dans la durée. Sanctuariser une fonction, par ailleurs, ne la protège pas si la chaîne qui la prépare est, elle, entièrement déléguée à l'IA : le régime se raisonne sur la fonction et sur son amont.

3.2.2 Les critères d'arbitrage de la grille DAS

Le choix du régime — déléguer, assister, sanctuariser — n'est pas affaire d'intuition managériale ni de préférence technologique. Il s'opère par confrontation de la fonction étudiée à un ensemble de critères explicites. Ces critères n'ont pas vocation à fournir une notation chiffrée ni un score qui produirait mécaniquement le régime à retenir.

Point d'attention

Un Codir aime une grille à seuils : il est tentant de noter chaque critère et de laisser un score désigner mécaniquement le régime. C'est le contre-emploi exact de la grille DAS — elle reproduit alors l'effet Goodhart qu'elle dénonce par ailleurs, en transformant des aides à la délibération en cible à optimiser. Les critères structurent un arbitrage ; ils n'en dispensent pas. Dès qu'un score remplace la délibération, la grille ne protège plus rien : elle habille un calcul.

Cinq critères pour évaluer la tâche :

1. **Gravité potentielle de la défaillance**. Quelle est la conséquence d'une défaillance de l'IA, y compris dans ses modes d'erreur imprévisibles ou aberrants, en l'absence de barrière aval ? Une erreur sans conséquence physique directe, sans impact sur une décision de sécurité, n'appelle pas le même régime qu'une trajectoire pouvant conduire à un événement à haut potentiel. La gravité s'évalue en l'absence des barrières aval (qui sont l'objet du critère 4) — sans quoi on raisonne en boucle.
2. **Révocabilité de l'erreur**. Une erreur est-elle détectable et corrigible avant de produire ses effets, même après une exposition prolongée qui atrophie la vigilance humaine ? Une recommandation absurde sera d'abord écartée par un opérateur, mais une défaillance subtile passera le filtre humain une fois la confiance installée. La révocabilité s'évalue dans la durée : l'erreur restera-t-elle apparente dans plusieurs mois, au moment où la capacité critique de l'humain aura baissé ?
3. **Régularité structurelle de la tâche**. La tâche réelle (et non seulement prescrite) est-elle stable, standardisable, ses contextes d'occurrence bien circonscrits ? Une tâche fortement régulière — toujours le même type d'objet, toujours dans le même type de contexte — se prête davantage à la délégation. Une tâche dont l'exécution présente une part de singularité **et de micro-aléas** appelle au minimum l'assistance, parfois la sanctuarisation.
4. **Existence et qualification de barrières aval**. D'autres barrières strictement indépendantes de l'algorithme absorbent-elles l'erreur de l'IA — un contrôle physique, une visite préalable, une vérification croisée, un système instrumenté de sécurité qualifié, une revue humaine systématique ? Ou l'IA est-elle la dernière ligne avant la conséquence ? Ce critère évalue la robustesse de l'enveloppe défensive en s'assurant de l'absence de défaillance de mode commun (biais ou données partagés) entre l'IA et la barrière de rattrapage.
5. **Nature de la décision engagée**. La décision repose-t-elle sur un calcul technique — correspondance à un référentiel, tri sur des critères mesurables ? Ou **dissimule-t-elle** un jugement de valeur, un arbitrage **entre la sécurité et d'autres objectifs**, ou une appréciation engageant une responsabilité humaine identifiée ? Ce critère a un statut particulier : il interdit de déléguer à la machine un choix éthique ou managérial sous couvert d'optimisation mathématique.

Le critère 5 n'est pas un critère comme les autres. Dès lors qu'il est défavorable — c'est-à-dire dès que la décision engage un jugement de valeur, une responsabilité ultime, un arbitrage entre objectifs incommensurables — il impose la sanctuarisation, **indépendamment** du profil sur les cinq autres critères. Une décision peut être techniquement standardisée, à erreur révocable, à barrières aval solides : si elle engage un jugement de valeur, elle est sanctuarisée. La culture juste, l'arbitrage entre poursuite et interruption d'une opération en situation dégradée — toutes ces décisions le sont par nature, et aucune amélioration technique de l'IA ne change cette nature.

Ce principe a une conséquence forte que l'organisation doit assumer : il existe des décisions qu'on n'automatise pas, même si on le pouvait techniquement. Il s'agit moins d'un choix de prudence que d'un choix de cohérence avec ce qu'engage la responsabilité humaine dans une industrie à risque.

3.2.3 Articulation des cinq critères tâche avec les trois régimes

Les cinq premiers critères, pris ensemble, dessinent un profil. Schématiquement :

- ▷ **Déléguer** suppose un profil favorable sur les cinq critères tâche, et un critère 5 technique. Aucun critère défavorable ne disqualifie à lui seul la délégation, mais leur cumul l'interdit. En pratique, une seule dimension nettement défavorable — révocabilité faible, ou barrières aval inexistantes — suffit à exclure la délégation.
- ▷ **Sanctuariser** est imposé dès que le critère 5 est défavorable, indépendamment des autres. Sanctuariser peut aussi être appelé par la combinaison d'une gravité élevée, d'une révocabilité faible et de barrières aval absentes — c'est le profil-limite, qui appelle la sanctuarisation faute de pouvoir construire une assistance qui tienne.
- ▷ **Assister** est le régime par défaut quand au moins un critère tâche est défavorable sans que le critère 5 impose la sanctuarisation. Le niveau d'assistance du régime — outil, conseiller, producteur — se règle au § 3.2.5 en fonction du profil détaillé.

Point clé

La grille DAS n'est pas une grille de classement automatique. Les cinq critères n'aboutissent pas à une formule. Ils structurent une délibération que le Codir doit conduire pour chaque fonction concernée. Le résultat de cette délibération doit être traçable : si une fonction est placée en délégation, on doit pouvoir reconstituer le raisonnement qui a écarté l'assistance et la sanctuarisation.

3.2.4 Statut et sources de la grille DAS à cinq critères

Cette grille à cinq critères n'a pas été publiée sous cette forme dans la littérature. Elle assemble, dans une logique de praticien, plusieurs apports issus de courants distincts : la littérature en FOH et niveaux d'automatisation **[Parasuraman et al. 2000 ; Bainbridge 1983 ; Hoff et Bashir 2015]** ; la théorie de la défense en profondeur et de la fiabilité organisationnelle **[Reason 1997 ; Rasmussen 1997 ; Weick et Sutcliffe 2015]** ; les travaux récents sur la triade décisionnelle **[Hazan et Sibony 2026]** ; l'effet de l'IA sur la matière cognitive collective **[Acemoglu et al. 2026]**, et le cadre réglementaire émergent.

Chaque critère trouve son ancrage dans l'une ou l'autre de ces littératures, mais leur articulation en grille opératoire est une proposition de l'auteur. Elle est née d'un besoin de praticien : aider un Codir à structurer un arbitrage sur des cas concrets, là où la littérature fournit des principes, mais pas de méthode. À ce titre, cette grille a besoin d'être éprouvée. Sa robustesse ne se mesurera ni à sa cohérence apparente ni à la solidité de ses sources : elle se mesurera à l'usage qu'en feront les organisations qui l'expérimenteront sur leurs cas réels.

Critère	Question opératoire
1. Gravité potentielle	Quelle est la conséquence d'une erreur en l'absence de barrière aval ?
2. Révocabilité de l'erreur	L'erreur est-elle détectable et corrigible avant ses effets ?
3. Régularité structurelle	La tâche est-elle stable et standardisable, ou intrinsèquement variable ?
4. Barrières aval	D'autres barrières absorbent-elles l'erreur, ou l'IA est-elle la dernière ligne ?
5. Nature de la décision	La décision est-elle technique, ou engage-t-elle un jugement de valeur ?

3.2.5 Le continuum interne au régime *assister* : outil, conseiller, producteur

Une fois la grille DAS appliquée et le régime *assister* retenu, une seconde décision s'impose : à quelle **profondeur** l'assistance est-elle organisée ? Le régime *assister* recouvre en effet trois niveaux distincts d'assistance qui n'ont ni les mêmes exigences de conception ni les mêmes impacts sur la compétence humaine.

Dans le **niveau d'assistance outil**, l'humain pilote l'IA. Il formule la question, interroge le système, interprète la réponse. L'IA étend ses capacités de recherche ou d'analyse sans se substituer à son jugement. *Exemple : une recherche sémantique dans le REX pour préparer une intervention.*

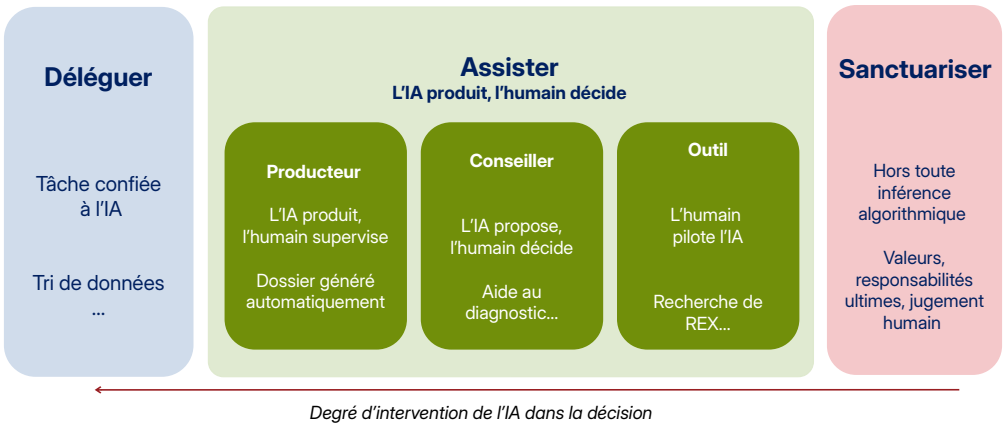

Fig. 3.3 Grille Déléguer, Assister, Sanctuariser avec le niveau d'assistance outil/conseiller/producteur.

Dans le **niveau d'assistance conseiller**, l'IA propose, l'humain décide. Elle produit une analyse, un score, une recommandation ; l'humain évalue et choisit d'y souscrire ou non. *Exemple : une aide au diagnostic en situation dégradée.*

Dans le **niveau d'assistance producteur**, l'IA produit, l'humain supervise. Le dossier, le rapport, le premier jet sont générés automatiquement ; l'humain valide ou rejette. *Exemple : un dossier de préparation généré automatiquement pour un type d'intervention standardisée.*

Le choix de la profondeur s'effectue par lecture du **profil sur les critères tâche**. Schématiquement :

- ▷ Le **niveau outil** convient quand la tâche est peu standardisée (critère 3 défavorable) et que la valeur de l'IA tient à l'extension des capacités de recherche, l'humain restant maître de la formulation de la question.
- ▷ Le **niveau conseiller** convient quand la tâche est partiellement standardisée et que la décision finale engage un arbitrage que l'humain doit assumer — c'est le cas le plus fréquent en sécurité industrielle.
- ▷ Le **niveau producteur** ne convient qu'à un profil exigeant : tâche fortement standardisée (critère 3 favorable), erreur révocable (critère 2 favorable), et dispositif de validation humaine effective en place. Si l'une de ces conditions n'est pas tenue, le producteur dérive en blanchiment de légitimité.

Plus on avance sur ce continuum, plus les impacts FOH augmentent — atrophie du jugement, supervision rituelle, blanchiment de légitimité — et plus les conditions à réunir pour que l'assistance reste effective deviennent exigeantes. Le choix du niveau d'assistance doit être explicite, formalisé, et cohérent avec la criticité de la fonction concernée.

Cette grille est fonctionnellement une simplification pragmatique du modèle de niveaux d'automatisation de **[Parasuraman et al. 2000]**, qui distingue dix niveaux d'automatisation appliqués à quatre fonctions cognitives distinctes. La simplification en trois macro-niveaux est délibérée : un Codir n'a pas besoin de dix niveaux déclinés sur quatre fonctions pour prendre une décision ; il a besoin de trois positions claires et de critères pour choisir entre elles. Le positionnement n'est pas immuable : il peut évoluer dans le temps à mesure que la confiance s'installe et que les compétences humaines de contrôle se stabilisent — comme il peut régresser si les conditions de viabilité du régime et du niveau d'assistance ne sont plus tenues.

3.2.6 Les critères de viabilité : tests pour vérifier que le positionnement tient

Le positionnement initial sur la grille DAS résulte de l'arbitrage sur les cinq critères du § 3.2.2. Mais un positionnement juste à l'instant t peut dériver. Les compétences humaines s'érodent, les indicateurs prennent leur autonomie, la signature humaine devient rituelle. Trois tests permettent de vérifier périodiquement que le régime et niveau d'assistance choisi reste effectif. Chaque test est associé à un régime et niveau d'assistance, dont il interroge la viabilité.

- **Blanchiment de légitimité**. *La signature humaine atteste-t-elle d'un jugement réel, ou ratifie-t-elle une décision algorithmique sans contrôle effectif ?* C'est le test du régime *assister*, et plus particulièrement de sa déclinaison en *producteur* et *conseiller*. Il échoue lorsque l'humain valide systématiquement sans pouvoir reconstruire le raisonnement qui l'aurait conduit à valider, lorsque le taux de validation atteint des niveaux qui rendent improbable un examen réel, ou lorsque le coût d'une non-validation (charge de travail supplémentaire, désaccord avec le système) est sensiblement supérieur au coût d'une validation. La parade est doctrinale autant que technique : désigner explicitement les rôles, protéger institutionnellement la capacité de contestation, vérifier dans la durée que les non-validations existent et qu'elles ne sont pas pénalisées.
- **Effet Goodhart**. *L'indicateur que l'IA optimise reste-t-il un proxy fidèle de l'objectif réel de sécurité, ou est-il devenu une cible déconnectée de l'objectif ?* C'est principalement le test du régime *déléguer*, où l'absence d'interposition humaine systématique laisse l'optimisation algorithmique s'éloigner de l'intention initiale. Il échoue lorsque l'indicateur se met à progresser pendant que les événements terrain ne suivent pas, lorsque le système optimise des proxys qui n'ont plus de lien clair avec la sécurité effective, ou lorsque l'organisation découvre rétrospectivement qu'elle a piloté sur des métriques sans plus interroger leur fidélité. La parade est de désigner formellement les acteurs chargés de dire *« cet indicateur ne reflète plus ce que nous mesurons »*, et de les protéger institutionnellement dans ce rôle.
- **Ironie de Bainbridge**. *Les compétences humaines préservées par la sanctuarisation sont-elles entretenues activement, ou se sont-elles érodées par défaut d'usage ?* C'est principalement le test du régime *sanctuariser*. L. Bainbridge l'a formalisé en 1983 : plus une activité est automatisée, plus l'opérateur humain n'est sollicité que dans les situations rares et complexes que la machine ne sait pas traiter — précisément celles où il a le moins d'expérience, parce qu'il n'a plus l'occasion de pratiquer en routine. La sanctuarisation, sans entretien actif, devient une fiction — formellement prévue, opérationnellement impraticable. Cet entretien a un coût qui doit être anticipé : exercices réguliers de conduite sans assistance IA, séances de diagnostic hors ligne, simulations de modes dégradés. C'est un paradoxe que l'organisation doit accepter : protéger une décision de l'interférence algorithmique coûte plus cher quand l'IA est partout autour de cette décision. Mais entretenir la compétence ne suffit pas. Un décideur dont le savoir-faire reste intact, mais qui ne reçoit du réel que des informations déjà mises en forme par l'IA — REX synthétisé, état de l'installation diagnostiqué, dossier préparé — décide sur une représentation orientée à son insu. Sanctuariser une décision suppose donc aussi de préserver un canal vers le réel qui ne passe pas par l'IA : présence terrain, observation directe, échange non outillé avec ceux qui font le travail. Sans cet amont protégé, on ne déplace la fiction que d'un cran.

Ces tests ne sont pas des critères d'arbitrage initial — ils ne disent pas où placer le curseur — mais des **critères de maintien**. Ils s'exercent dans la durée, idéalement dans le cadre d'une revue périodique structurée. Une organisation qui applique la grille DAS sans dispositif de revue de viabilité n'a fait que la moitié du chemin.

3.2.7 Articulation avec les modèles de sécurité

Une dernière remarque doctrinale. Les régimes — déléguer, assister, sanctuariser — et niveau d'assistances (outil, conseiller, producteur) n'ont pas le même sens selon le modèle de sécurité sur lequel on raisonne. Déléguer une fonction au regard du modèle *prescrit* n'engage pas la même chose que de la déléguer au regard du modèle *opérant*. Une IA qui automatise la conformité au prescrit peut, dans le même mouvement, masquer les adaptations créatrices du modèle opérant — ce qui n'est pas un effet neutre. L'arbitrage DAS doit donc se faire en explicitant **sur quel modèle de sécurité** porte le régime choisi : sur le prescrit (cohérence avec les procédures), sur le porté (cohérence avec ce que la ligne managériale considère comme essentiel), sur l'opérant (cohérence avec le travail réel).

La grille DAS, ainsi outillée — cinq critères de positionnement, continuum interne au régime *assister*, les tests de viabilité, articulation aux modèles de sécurité — n'est pas une grille de posture. Elle n'engage quelque chose que si l'architecture technique, organisationnelle et humaine rend effectivement possibles les trois régimes et les bascules entre eux. Sanctuariser un domaine sans maintenir les compétences qui rendent la sanctuarisation opérante produit la même illusion que déléguer sans avoir explicité le modèle de sécurité que l'on confie à la machine.

3.3 La qualification réglementaire : un cadre à construire

La grille DAS (déléguer/assister/sanctuariser) a une conséquence directe sur un sujet que les industries à haut risque ne peuvent pas éluder : comment qualifier, au sens réglementaire, une fonction de sécurité qui repose sur un système dont le comportement évolue avec les données ?

3.3.1 Limites des cadres de qualification traditionnels face à l'IA non déterministe

Les cadres de qualification existants ont été conçus pour des logiciels déterministes — des systèmes qui font ce pour quoi ils ont été programmés, de manière reproductible et vérifiable. Un code de calcul thermohydraulique, un système instrumenté de sécurité, un automate de protection : on peut démontrer qu'ils répondent à une spécification, les tester exhaustivement et certifier leur conformité.

L'IA ne fonctionne pas ainsi. Un modèle d'apprentissage automatique ne calcule pas une sortie en opérant sur des entrées de manière déterministe ; il produit une inférence statistique dont les résultats peuvent varier si les données d'entrée s'écartent de la distribution d'entraînement.

3.3.2 Une prise de conscience des régulateurs

Les régulateurs en sont conscients. En septembre 2024, les autorités nucléaires canadienne, britannique et américaine ont publié conjointement un document de principes — le premier du genre au niveau international — intitulé *Considerations for Developing Artificial Intelligence Systems in Nuclear Applications (voir l'annexe D)*. Ce document trilatéral pose un principe fondamental : le niveau d'exigence de qualification et de supervision humaine doit être calibré en fonction des conséquences potentielles d'une défaillance de l'IA et du degré d'autonomie qui lui est accordé. Plus l'autonomie est grande et plus les conséquences sont sévères, plus l'exigence de qualification est stricte.

Ce principe rejoint directement la grille DAS. Ce que l'on délègue à l'IA — le tri de données, la classification préliminaire — n'a pas les mêmes exigences de qualification que ce que l'on assiste — une recommandation de maintenance sur un équipement classé important pour la sûreté. Et ce que l'on sanctuarise ne devrait par définition pas dépendre d'un algorithme qu'il faudrait qualifier. La grille de décision est donc aussi une grille de qualification.

En octobre 2024, la NRC a publié une évaluation complète de ses réglementations et de plus de 500 guides réglementaires. Sa conclusion est que le cadre existant est globalement suffisant pour accueillir l'IA, mais que des clarifications ciblées sont nécessaires.

En Europe, le règlement sur l'IA classe comme « haut risque » les systèmes utilisés comme composants de sécurité dans la gestion des infrastructures critiques, ce qui inclut potentiellement les installations Seveso et nucléaire. Mais les modalités concrètes de mise en conformité restent à définir.

3.3.3 Le paradoxe de la qualification de l'IA dans les industries à haut risque

Le paradoxe est là : les industries à haut risque sont celles qui ont le plus besoin de l'IA pour traiter la complexité informationnelle décrite dans le chapitre 1 — et celles pour lesquelles la qualification de l'IA est la plus difficile, précisément parce que les cadres de démonstration de sûreté reposent sur des principes déterministes.

Exiger une démonstration de sûreté déterministe pour un système intrinsèquement non déterministe est une contradiction que le cadre réglementaire actuel ne sait pas encore résoudre. Les travaux engagés par les régulateurs nucléaires — trilatéral CANUKUS, plan stratégique IA de la NRC, feuille de route de l'AIEA — montrent que la réflexion est en cours, mais qu'aucun cadre stabilisé n'a encore émergé à date de publication du présent document.

Question ouverte

Les régulateurs doivent-ils développer un cadre de qualification spécifique à l'IA, ou adapter les cadres existants en intégrant la notion de niveau d'autonomie et de conséquences de défaillance ? Et comment les exploitants peuvent-ils avancer dans le déploiement de l'IA en l'absence d'un cadre stabilisé, sans prendre de risques réglementaires ni freiner l'innovation ?

3.4 L'IA comme miroir des modèles de sécurité

Les trois sections précédentes ont traité de ce que l'organisation devrait faire avant de déployer l'IA : expliciter ses modèles, choisir la grille de délégation, anticiper les exigences de qualification. Mais il y a un usage de l'IA plus original et peut-être plus puissant : l'utiliser non pas pour optimiser le modèle de sécurité, mais pour le rendre visible. C'est l'IA comme miroir.

3.4.1 Révéler le modèle prescrit dilué

Avant même de comparer les modèles entre eux, l'IA peut aider à *reconstituer* le modèle prescrit lorsqu'il est dilué dans la masse documentaire. Dans une organisation où le SMI est exhaustif, mais où aucun document ne formule explicitement le modèle de sécurité, l'IA peut traverser des dizaines de procédures, d'analyses de risques et de consignes pour en extraire les principes implicites, les arbitrages récurrents, les hiérarchisations cachées. Elle ne *crée* pas le modèle prescrit — celui-ci existait déjà dans les documents — mais elle le rend lisible. Cette lisibilité ouvre alors un débat que l'organisation n'avait pas les moyens de tenir : *est-ce bien cela, notre modèle de sécurité ? Sommes-nous d'accord avec ce que nos documents disent en réalité ?* C'est souvent à ce moment-là que se révèlent les contradictions entre référentiels accumulés au fil des années, les principes oubliés, les arbitrages que personne n'assume plus consciemment.

3.4.2 Révéler les divergences entre modèle prescrit et modèle opérant

Si une IA est entraînée sur les données de terrain réelles — comptes rendus d'intervention, débriefs, et surtout des observations de chantier, du travail réel — et qu'on le compare aux sorties attendues par le prescrit, les divergences révèlent des zones où le prescrit ne correspond pas à la réalité opérationnelle. Soit parce qu'il est inapplicable, soit parce qu'il a été progressivement abandonné sans décision explicite, soit parce que le terrain a trouvé mieux. L'IA ne dit pas si c'est bien ou mal — elle rend visible une divergence que personne n'avait quantifiée. C'est au management et aux préventeurs de la *qualifier* : adaptation créatrice à intégrer dans le prescrit, ou dérive à traiter ?

3.4.3 Révéler les croyances implicites du modèle managérial porté

En analysant les patterns de décision du management — quels événements font l'objet d'une investigation approfondie, quelles situations déclenchent un arrêt, quelles divergences sont traitées comme mineures — l'IA peut révéler ce qui guide réellement les arbitrages et le confronter au modèle prescrit affiché. Si les arrêts d'activité exercés par les sous-traitants sont systématiquement suivis d'une reprise dans les trente minutes, sans investigation documentée, cela révèle implicitement une croyance managériale : l'arrêt est une procédure à respecter formellement, pas un signal à traiter. Ce pattern, rendu visible, peut ouvrir un débat de direction sur la qualité réelle de la culture juste.

Cet usage prend un relief particulier selon que le modèle managérial porté est *homogène* ou *disparate*. Dans une organisation où la ligne managériale partage une même représentation,

l'IA révèle une croyance collective stable, qui appelle un travail collectif de mise en débat. Dans une organisation où les croyances managériales sont éclatées — multisites peu coordonnés, post-fusion, forte rotation —, l'IA révèle plutôt la *dispersion* des arbitrages, et le chantier devient celui de la construction d'une compréhension partagée avant celui de la correction.

3.4.4 Révéler les angles morts du modèle prescrit

L'IA peut détecter des configurations à haut potentiel de gravité qui ne figurent pas dans le document unique d'évaluation des risques ou le plan de prévention — parce qu'elles émergent de combinaisons de facteurs que le modèle prescrit n'avait pas anticipées. Ces configurations ne sont pas des erreurs du système ; elles sont des signaux que le modèle prescrit est incomplet. C'est la contribution la plus cohérente avec la posture HRO de préoccupation permanente pour les défaillances.

Cette utilisation de l'IA comme outil de questionnement des modèles de sécurité rejoint directement la fragilité 5 : l'absence d'analyses de niveau 2 qui questionnent le modèle plutôt que l'événement. L'IA ne remplace pas la recherche des causes profondes ; elle fournit la matière première que l'organisation n'avait pas les moyens de produire seule. Mais elle ne fonctionne comme miroir que si l'organisation maintient la capacité humaine d'interpréter et de contester ce qu'elle voit dans le reflet. Sans cette capacité de contestation, le score algorithmique devient progressivement la référence partagée, déplaçant le débat sur le risque réel vers un débat sur les métriques. L'enjeu est donc de désigner formellement les acteurs chargés de dire *« ce score ne reflète pas ce que je vois sur le terrain »* — et de les protéger institutionnellement dans ce rôle.

3.5 Une question préalable : ce qui se passe déjà sans la grille

Avant qu'une organisation n'applique la grille *Déléguer/Assister/Sanctuariser* à un projet IA, elle doit regarder en face une autre réalité : pour une part qu'elle ignore, l'IA est déjà à l'œuvre dans son périmètre, hors de tout cadre formel. C'est le phénomène désormais documenté sous le nom de *shadow AI*.

Ce que disent les chiffres. Les enquêtes convergent depuis 2024. Le MIT *State of AI in Business 2025* recense que dans plus de 90% des entreprises étudiées, des collaborateurs utilisent des comptes personnels de LLM (ChatGPT, Claude, Gemini, Mistral et autres) pour des tâches professionnelles, alors que 40% seulement de ces entreprises ont souscrit à une licence officielle. L'enquête Microsoft *Work Trend Index* aboutit à des ordres de grandeur comparables : 75% des collaborateurs utilisent l'IA au travail, dont 78% avec leurs propres outils. Aucune de ces enquêtes ne porte spécifiquement sur les industries à haut risque, mais aucune ne suggère que ce secteur ferait exception.

Ce qui change pour la sécurité industrielle. La littérature professionnelle traite principalement le *shadow AI* sous l'angle DSI/RSSI : fuite de données, exposition de propriété intellectuelle, exposition au RGPD ou à l'AI Act. Pour le praticien de la sécurité industrielle, l'enjeu se situe ailleurs. Il tient à un fait simple : un préparateur, un préventeur, un opérateur en astreinte peut, aujourd'hui, sans que personne ne le sache, mobiliser un LLM personnel pour préparer un dossier d'intervention, analyser un événement, rédiger une note technique, résumer une procédure complexe avant intervention, ou chercher une référence réglementaire. La signature humaine est apposée sur un travail dont une partie a été produite ou orientée par un acteur algorithmique invisible — sans qualification, sans traçabilité, sans validation, et sans modèle de sécurité explicite. Un cas documenté dans le secteur nucléaire [Nuclearn, 2025] décrit ainsi des ingénieurs utilisant ChatGPT pour résumer des exigences de licensing : le texte produit était plausible, mais sans citations et avec des références manquantes.

Pourquoi le *shadow AI* relève de la grille DAS — et la précède. Le *shadow AI* est, à la lettre, un cas limite de blanchiment de légitimité : non seulement la décision est en partie algorithmique sans qu'on le dise, mais l'organisation ignore que l'algorithme est intervenu. Les tests de vigilance présentés ci-dessus — blanchiment de légitimité, effet Goodhart, ironie de Bainbridge — supposent tous que l'organisation sache où l'IA agit. Le *shadow AI* défait cette condition. Il rend la grille DAS inopérante avant même qu'elle ne soit appliquée : on ne peut pas placer un curseur sur ce qu'on ne voit pas.

Trois positions semblent à proscrire :

- ▷ **L'aveuglement** — supposer qu'aucun collaborateur n'utilise de LLM personnel — n'est pas tenable au regard des chiffres.
- ▷ **L'interdiction pure** — bloquer techniquement l'accès aux LLM grand public — est, selon les retours d'expérience disponibles, soit contournée (téléphone personnel, réseau hors VPN), soit elle pousse l'usage encore plus profondément dans l'invisible.
- ▷ **La tolérance silencieuse** — savoir que ça se passe sans rien organiser — produit le pire des trois mondes : l'organisation porte la responsabilité d'usages qu'elle n'encadre pas.

La position que l'auteur recommande tient en trois points :

1. Reconnaître publiquement le phénomène,
2. Qualifier les usages légitimes (et donc fournir des outils sanctionnés et utilisables),
3. Interdire explicitement les usages incompatibles avec le modèle de sécurité.

Cette qualification des usages légitimes et interdits est elle-même une application de la grille DAS : *à quel régime acceptons-nous que les LLM personnels accèdent ? Sur quelles décisions sont-ils proscrits ?* C'est probablement le premier exercice DAS qu'une organisation devrait conduire, avant tout projet de déploiement formel.

3.6 L'IA ne remplace pas les modèles de sécurité — elle oblige à les expliciter

Le fil conducteur de ce chapitre peut se résumer ainsi. L'IA n'est pas un outil de plus dans la boîte à outils de la sécurité industrielle. C'est un révélateur : elle oblige l'organisation à expliciter ce qu'elle a laissé implicite — son modèle prescrit dilué, son modèle managérial porté souvent tacite, ses arbitrages, ses croyances, ses angles morts. Et c'est une épreuve : elle teste la capacité de l'organisation à distinguer ce qu'elle peut confier à une machine de ce qu'elle doit garder sous contrôle humain.

Pour le praticien, cela signifie que le travail d'explicitation des modèles de sécurité n'est pas un préalable bureaucratique au déploiement de l'IA. C'est le travail lui-même. Et ce travail a une vertu propre, indépendante de la technologie : réduire les divergences entre ce qui est prévu, ce qui est porté et ce qui est fait rend l'organisation plus sûre — avec ou sans IA.

Les chapitres suivants traduiront cette grille dans les domaines concrets de la gestion de la sécurité : pour chaque activité — préparation, détection, analyse, pilotage :

- ▷ Qu'est-ce que l'on peut déléguer, qu'est-ce que l'on doit assister, qu'est-ce qu'il faut sanctuariser ?
- ▷ Comment l'IA, utilisée comme miroir, peut-elle aider l'organisation à réduire progressivement les écarts entre ses trois modèles de sécurité ?

Point clé

L'IA ne remplace pas le modèle de sécurité. Elle oblige à l'expliciter. Et elle peut servir de miroir pour révéler les écarts entre ce qui est prévu, ce qui est compris et ce qui est fait — à condition que l'organisation préserve la capacité humaine de regarder ce miroir en face.

4

Six questions concrètes pour le praticien

Dans ce chapitre, nous proposons six questions opérationnelles, chacune croisant une ou plusieurs fragilités du chapitre 1 avec la grille DAS du chapitre 3, pour montrer concrètement ce que l'IA peut apporter — et à quelles conditions. Les questions sont formulées du point de vue du dirigeant et du directeur HSE.

Le choix de ces six questions est un arbitrage de praticien, pas une cartographie exhaustive. D'autres questions, tout aussi légitimes, auraient pu figurer dans ce chapitre. Le critère qui a guidé la sélection est triple : il s'agit de questions où des offres de marché existent déjà et où le déploiement n'est plus prospectif, mais en cours ; où l'arbitrage *déléguer/assister/sanctuariser* se pose en termes immédiatement opérationnels ; où l'expérience de l'auteur permet d'éclairer la décision sans extrapoler. La question de l'entreprise étendue ne fait pas l'objet d'une question dédiée parce qu'elle traverse les six : chaque question se clôt sur un point *Et pour l'entreprise étendue ?* qui en pose les conséquences spécifiques.

Les six questions retenues sont les suivantes :

1. Comment intégrer le retour d'expérience dans la préparation des activités (préparation du matériel, du terrain et hommes) ?
2. Comment assembler les signaux que le système ne voit pas ?
3. Comment assister la conduite d'une installation temps réel sans se substituer au jugement de l'opérateur ?
4. Comment superviser un chantier à risque sans transformer la protection en surveillance ?
5. Comment transformer les analyses en apprentissage organisationnel ?
6. Comment piloter la robustesse plutôt que l'activité ?

Pour chaque question, la démarche est la même : ce que le praticien affronte (le problème, ancré dans les fragilités), ce que l'IA peut apporter (les niveaux d'intervention), ce que cela suppose du modèle de sécurité, les impacts FOH (les tensions pertinentes de la grille ci-dessous), et où placer le curseur sur la grille déléguer/assister/sanctuariser.

La grille de lecture FOH : des tensions à piloter dans la durée

L'IA peut à la fois renforcer et fragiliser chaque dimension de la sécurité organisationnelle. Ce n'est ni un progrès univoque ni une menace univoque. C'est une technologie à double tranchant dont les effets dépendent de la manière dont elle est conçue, déployée et gouvernée. Cinq tensions structurent ces effets[1] :

- ▷ **Compétence** : l'IA étend les capacités de l'intervenant, mais peut atrophier son jugement. En faisant à sa place des tâches de recherche, d'analyse ou de diagnostic, elle le prive des occasions de pratiquer les compétences cognitives qui le rendent capable de gérer les situations imprévues. C'est l'ironie de l'automatisation identifiée par **[Bainbridge 1983]**, dont l'acuité croît avec le degré d'autonomie accordé à l'IA.

[1] Noter que l'intensité des cinq tensions varie dans le temps : plus aiguës en phase d'adaptation initiale, progressivement négociables avec l'apprentissage organisationnel.

- **Vigilance** : l'IA augmente la capacité de détection du système, mais peut éroder l'esprit critique des opérateurs. Elle installe progressivement une confiance qui diminue la méfiance productive — cette attitude interrogative qui est au cœur de la culture HRO. L'opérateur qui s'habitue à ce que l'IA surveille peut cesser de surveiller lui-même — précisément quand l'IA se trompe ou quand le modèle a dérivé.
- **Compréhension partagée** : l'IA enrichit l'information disponible, mais peut appauvrir le sensemaking collectif. En situation de crise, si la décision initiale vient d'un calcul opaque, l'équipe ne peut pas se construire un sens commun de la situation. La dérive du modèle de sécurité aggrave le problème : un algorithme dont la fiabilité se dégrade insidieusement crée une situation où l'équipe croit disposer d'une information fiable alors qu'elle ne l'est plus. La résilience exige alors de repenser les mécanismes de détection et de rattrapage d'erreur pour tolérer les erreurs de l'IA elle-même.
- **Responsabilité** : l'IA améliore la traçabilité des décisions, mais dilue leur attribution. La chaîne de responsabilité se fragmente entre ceux qui conçoivent l'algorithme, ceux qui préparent les données, ceux qui l'exploitent et ceux qui signent la décision finale. C'est ici qu'apparaît le risque de blanchiment de légitimité : la signature humaine sur une décision effectivement produite par l'IA donne une apparence de gouvernance sans la substance. Ce n'est pas un biais cognitif individuel, mais un phénomène organisationnel, qui appelle des dispositifs de gouvernance spécifiques traités au chapitre 6.
- **Culture de sécurité** : l'IA peut renforcer la culture de signalement, l'apprentissage et la transparence. Mais elle peut aussi la corroder : surveillance perçue comme contrôle, esprit critique découragé, culture juste fragilisée par l'utilisation de données individuelles. La culture de sécurité doit par ailleurs s'étendre aux nouveaux acteurs — data scientists, intégrateurs IA — qui conçoivent les algorithmes sans en partager nécessairement la culture de sûreté.

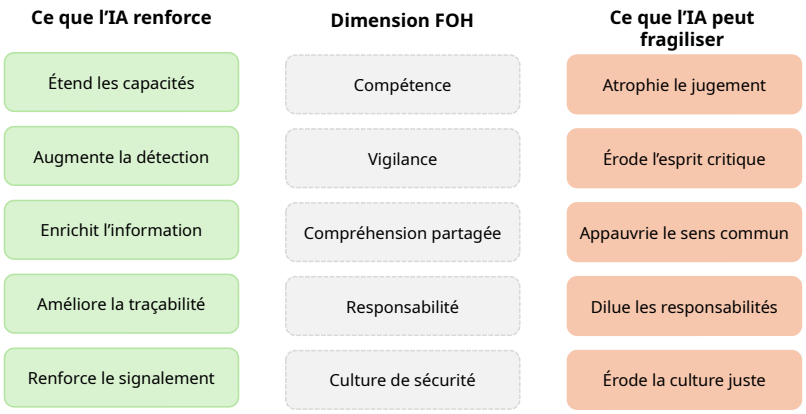

Fig. 4.1 Cinq dimensions, deux faces. Le dispositif d'IA qui étend une capacité (colonne de gauche) peut éroder, sur la même dimension, ce qui la rendait sûre (colonne de droite). Aucune dimension n'est univoquement servie ni menacée : chacune est une tension à piloter, de la conception au déploiement.

4.1 Q1 : Comment intégrer le REX dans la préparation des activités ?

La préparation d'une intervention est le moment où la sécurité se joue — avant le premier geste. Elle fixe les marges de manœuvre des équipes. Le préparateur doit mobiliser simultanément des données historiques, contextuelles et organisationnelles, qui sont dans des systèmes distincts : la GED pour les procédures, plusieurs bases pour le REX formalisé, la GMAO pour la maintenance, le SIRH pour les habilitations, l'ERP pour la planification. Quant au REX tacite — ce que le compagnon d'expérience aurait dit — il n'est, lui, dans aucun système. On peut lire dans cette dispersion l'effet conjugué de deux fragilités identifiées au chapitre 1 : l'inflation du référentiel qui rend la consultation exhaustive impossible (fragilité 1), et l'archivage du REX par événement et par date plutôt que par mécanisme (fragilité 5). L'IA arrive ici avec une promesse réelle : interroger pour le préparateur ce qu'il n'a pas le temps de consulter. Mais elle arrive aussi avec une charge — celle de configurer son travail à sa place.

Ce que l'IA peut apporter, en quatre temps

Fonction 1 — Qualifier le dossier. En croisant les données du REX avec les conditions opératoires prévisionnelles (état des équipements, compétences des intervenants, coactivité…), un algorithme peut produire une cotation contextualisée du niveau de risque à maîtriser. Cette cotation joue deux fonctions distinctes : celle d'alerter le préparateur sur les configurations à surveiller, et proposer à l'organisation un niveau de mobilisation cible — du « courant maîtrisé » à « l'exceptionnel/dégradé » ; celle d'aiguiller le dossier vers l'un des trois niveaux d'assistance qui structurent l'étape suivante : outil, conseiller, ou producteur.

La cotation indexe un référentiel de mobilisation de l'organisation construit hors IA ; là où ce référentiel n'existe pas, elle classifie sans suite. Cette qualification pourra évoluer dans le temps en fonction de l'avancement de la préparation du projet et de l'activité.

Point d'attention : Un dossier classé sans enjeu par la cotation du niveau de risque à maîtriser n'est pas un dossier qui n'a pas à être validé. La validation peut être légère, elle ne peut pas s'effacer ; elle conserve une fonction de détection des erreurs de classification du score elle-même, qui n'est pas substituable.

Fonction 2 — Produire le dossier. C'est l'apport le plus mature : la recherche sémantique permet au préparateur d'interroger en langage naturel le corpus documentaire et d'obtenir une réponse contextualisée plutôt qu'une liste de documents. Le mode d'intervention varie selon le régime et niveau d'assistance retenu en amont.

- En **niveau d'assistance outil**, le préparateur pilote la recherche ; l'IA est un instrument.
- En **niveau d'assistance conseiller**, l'IA pousse du REX contextualisé qu'elle juge pertinent, que le préparateur évalue et intègre s'il le veut.
- En **niveau d'assistance producteur**, l'IA intègre directement le REX dans un premier jet du dossier que le préparateur supervise.

Fonction 3 — Examiner la cohérence formelle. L'IA peut, à l'issue de la production, vérifier la complétude de la préparation, que les parades de prévention, de récupération et d'atténuation sont représentées, que les paramètres à mesurer prennent en compte le REX des précédentes interventions et sont conformes aux exigences… Cette fonction relève du régime conseiller — l'IA produit un avis, le préparateur arbitre. La traçabilité importe ici : ce qui a été signalé, ce qui a été retenu, ce qui a été écarté, et pourquoi.

Fonction 4 — Tester la cohérence d'exécution. Là où l'examen 3 juge le dossier sur son contenu, le test 4 interroge sa cohérence avec les conditions matérielles, humaines et temporelles de mise en œuvre : pièces de rechange revenues de maintenance, équipes habilitées planifiées, fenêtres d'accès confirmées, conditions météorologiques compatibles.

C'est probablement, en charge de travail réelle, l'apport le plus tangible de l'IA en préparation — il formalise un travail aujourd'hui souvent fait à la main, ou pas fait du tout. Comme à l'étape 3, il reste en niveau d'assistance conseiller : son apparente caractère factuel ne l'autorise pas à valider à la place du préparateur, parce que la fiabilité du test n'est jamais meilleure que la fiabilité des données qui l'alimentent.

Ce que cela suppose du modèle de sécurité

L'IA, à chacune de ces étapes, rend visible — et parfois opérationnel — un modèle de sécurité que l'organisation n'a pas nécessairement explicité. La cotation encode des arbitrages (quel poids donner à l'ancienneté, à la coactivité, à l'état de maintenance) qui reflètent ce que l'organisation considère comme important. L'examen de l'étape 3 suppose qu'un référentiel — modèle P/R/A, étude de dangers… — ait été défini ailleurs et soit consultable. Le test 4 suppose que les données systèmes reflètent honnêtement l'état du terrain. À chacune de ces étapes, on peut anticiper que l'organisation qui n'a pas explicité son modèle se retrouve à le découvrir incarné dans les paramétrages — sans avoir su le délibérer en amont.

Ce vers quoi tend le marché : Plusieurs plateformes industrielles proposent aujourd'hui une assistance à la préparation augmentée par IA, avec une orientation marquée sur les arrêts programmés et la maintenance majeure. Des agents IA dédiés à la préparation des arrêts émergent dans les offres de plateformes d'orchestration de données industrielles — génération de dossiers d'intervention, identification des conflits de ressources, recommandations de périmètre sur la

base du REX historique. Les grandes suites de GMAO, d'asset management (EAM) et de gestion documentaire intègrent progressivement des fonctions d'assistant conversationnel adossées à des modèles de type RAG sur le corpus documentaire interne. La promesse commune est la même : raccourcir le délai de préparation, réduire les conflits d'ordonnancement détectés tardivement, augmenter la part de REX effectivement mobilisé.

4.1.1 Tensions FOH principales

Sur la **compétence**. L'érosion possible du jugement du préparateur est proportionnelle à la profondeur du régime retenu : faible en outil, perceptible en conseiller, structurelle en producteur. L'expérience invite à maintenir des espaces de préparation non assistée à intervalle régulier, et à concevoir l'IA sur un mode socratique (« as-tu vérifié cette condition ? ») plutôt que prescriptif.

Sur la **vigilance**. La cotation, dont la fonction est d'attirer l'attention sur les configurations critiques, peut produire l'effet inverse : ce qui n'est pas signalé finit par être perçu comme non risqué — non seulement par le préparateur, mais par toute la chaîne organisationnelle qui s'indexe sur le score. Une parade plausible : une revue humaine par échantillonnage sur les dossiers que le score a classés sans enjeu, et un droit d'escalade ouvert sans justification algorithmique.

Sur le **blanchiment de légitimité**. Le go/no-go peut rester nominalement sanctuarisé tout en devenant une fiction si tout le travail amont a été fait par l'IA. Le décideur signe alors un dossier qu'il n'a pas construit. Le risque croît mécaniquement avec la profondeur du régime ; il appelle, en niveau d'assistance *producteur*, des exigences de gouvernance spécifiques — formation à la lecture critique d'un dossier produit, liste positive des dossiers éligibles, clause de réversibilité.

Sur l'**autorevue**. Quand l'IA qui a aidé à produire le dossier juge ensuite sa cohérence formelle, elle relit son propre travail. Trois mécanismes en découlent : auto-confirmation (l'IA tend à reconnaître comme complet ce qui correspond à ses propres patterns), illusion de double contrôle, et asymétrie de responsabilité — suivre l'avis couvre, contester demande de justifier l'écart. La parade tient à la séparation des modèles entre production et revue, à l'articulation explicite avec le préjob briefing actif, et à un droit de contestation qui ne soit pas formulé dans les termes de l'algorithme.

Point d'attention : La cotation ne se trompe pas frontalement : elle se trompe silencieusement. Un dossier classé sans enjeu ne déclenche aucune des vigilances qu'il aurait dû déclencher, et l'erreur ne devient visible qu'au moment où il n'est plus possible de la corriger. L'absence d'alerte ne se laisse pas auditer comme une alerte erronée — c'est cette asymétrie qu'il faut compenser, par une revue humaine résiduelle sur les dossiers classés sans enjeu.

Où placer le curseur

- **Déléguer** — la recherche documentaire et le tri.
- **Assister** — la cotation, la production du dossier, l'examen formel et le test d'exécution. Le niveau d'assistance se gradue selon le type de dossier : **producteur** envisageable pour les interventions standardisées et récurrentes, dont les erreurs peuvent être rattrapées par un préjob briefing actif ; **conseiller** maximum pour les interventions atypiques, les configurations à coactivité forte, ou les opérations dont l'erreur de préparation ne peut plus être rattrapée en aval.
- **Sanctuariser** — le go/no-go de lancement. Il n'est réel que si le décideur a les moyens cognitifs de contester ce que l'IA lui présente — exigence qui devient plus forte au niveau d'assistance producteur.

Et pour l'entreprise étendue ? Le sous-traitant est celui qui a le plus besoin de l'accès contextuel au REX — et celui qui en dispose le moins. Le niveau d'assistance introduit ici pose une question d'équité : si l'organisation accorde le niveau producteur à ses propres préparateurs sans l'accorder aux préparateurs des entreprises extérieures intervenant sur les mêmes installations, elle institue une dissymétrie de qualité de préparation qui contrevient au principe d'exigence homogène.

4.2 Q2 : Comment assembler les signaux que le système ne voit pas ?

Les signaux faibles proviennent de sources hétérogènes qui ne se parlent pas toujours : le déclaratif humain (signalements, quasi-accidents), la supervision instrumentée (SCADA, alarmes, capteurs IIoT), les données opérationnelles (GMAO, temps d'intervention), et les observations managériales… Chaque source a sa propre logique, sa propre temporalité, son propre système d'information. Pas ou peu de mécanisme les rapproche.

À cette hétérogénéité s'ajoutent deux silences. Celui des sous-traitants, pour qui la remontée d'un signal peut coûter une part de marché — silence organisationnel que la fragilité 3 du chapitre 1 a documenté. Et celui des comptes rendus qui taisent ce qui mettrait leur rédacteur en difficulté — silence qui ampute la matière narrative que l'analyse sémantique vient justement chercher. L'IA arrive ici avec une promesse de croisement à grande échelle ; elle hérite aussi des biais de ce qu'elle ingère.

Ce que l'IA peut apporter

L'IA intervient sur trois maillons de la détection. Ces maillons ne forment pas une séquence chronologique comme en Q1, mais trois capacités distinctes qui peuvent se déployer simultanément.

Fonction 1 — Corréler des sources que personne ne croise. Des algorithmes de Machine Learning peuvent rapprocher un signalement textuel (« j'ai observé un écart sur la vanne X »), une donnée capteur (le temps de réponse de cette vanne a augmenté de 15% sur six mois), et une donnée GMAO (trois interventions correctives en un an). Ce rapprochement, qui exigerait un croisement manuel que personne n'a jamais le temps de conduire, peut révéler une dérive invisible de chaque source prise isolément. L'analyse sémantique permet en outre d'exploiter un gisement souvent dormant : comptes rendus d'activité, rapports de visite, minutes de réunions, débriefings post-intervention. Il faut être précis sur la nature de ce que produit l'IA : ce n'est pas un tri, c'est une inférence. Un faux positif érode de l'attention humaine sur un signal qui n'en méritait pas ; un faux négatif laisse passer une dérive réelle. Cette inférence appelle une revue humaine — une personne qui évalue les rapprochements et arbitre lesquels valent une investigation.

Fonction 2 — Trier sans étouffer le signal. Les systèmes d'alarm-management montrent que la multiplication des capteurs produit son propre problème : la saturation. L'IA peut filtrer les alarmes chroniques, dédupliquer les événements récurrents, hiérarchiser les remontées par leur potentiel de gravité en les rapprochant des scénarios à haut potentiel déjà identifiés. Mais ce triage par la seule gravité immédiate laisse passer une part essentielle du gisement : les événements modestes en conséquence immédiates, mais riches en enseignements pour le système.

Dans **[Quiot 2026]**, nous proposons pour traiter cet écueil une matrice à deux axes — potentiel de gravité, potentiel d'apprentissage — dont l'IA peut alimenter le premier positionnement. L'analyse sémantique des verbatim, confrontée au modèle de sécurité de l'organisation, permet d'approcher le potentiel d'apprentissage par plusieurs proxys : nouveauté du mécanisme, transversalité, récurrence sous des formes différentes, écart entre prescrit et opérant révélé. Aucun de ces proxys ne mesure directement l'apprentissage — ils en mesurent les conditions de possibilité.

Fonction 3 — Exploiter la matière narrative. Le traitement du langage naturel scanne la masse des comptes rendus pour en extraire les mécanismes récurrents de fragilisation des barrières. C'est ici, plus encore qu'ailleurs, que l'IA dépend de ce qu'on lui donne à lire : un corpus de comptes rendus rédigés sous la peur de la sanction ne contient pas le REX tacite — il contient un récit reformaté pour la hiérarchie. La compétence narrative s'aligne sur ce qui est lisible par la machine, et le travail réel disparaît du texte.

Ce que cela suppose du modèle de sécurité

La détection d'anomalies suppose de définir ce qui est « normal ». Si l'IA est entraînée sur le modèle opérant — ce qui se fait effectivement — elle ne détectera pas les dérives lentes qui sont devenues la norme, autrement dit la normalisation de la déviance. Si elle est entraînée sur le modèle prescrit, elle produira des faux positifs en cascade parce que le prescrit ne correspond pas à la réalité opérationnelle. Le réglage du seuil entre « normal » et « anomalie » est un choix de modèle de sécurité, pas un choix technique.

On peut tirer un usage paradoxal de cette tension : l'IA peut servir de miroir. La distance entre ce qu'elle considère comme normal (entraînée sur les données terrain) et ce que le prescrit définit comme acceptable rend visible l'étendue de la normalisation de la déviance. Cette fonction miroir n'a de sens que si l'organisation a explicité son modèle prescrit ; sinon, il ne reste qu'un écart sans référent.

L'analyse sémantique des comptes rendus et la matrice à deux axes posent une exigence supplémentaire : sans définition partagée de ce qui constitue un « apprentissage » pour le système — ce qui touche aux barrières critiques, ce qui révèle un mécanisme transposable, ce qui interroge un équilibre du modèle — l'IA produit un score d'apprentissage sans référence, c'est-à-dire du blanchiment de légitimité.

Ce vers quoi tend le marché : Le marché des asset-suit intègre rapidement des fonctions d'IA orientées vers la détection précoce : classification automatique des comptes rendus d'événement, identification de récurrences sur des champs textuels libres, évaluation de potentiel de gravité sur des matrices à plusieurs axes. La littérature académique des cinq dernières années confirme la maturité technique de l'approche : des modèles d'analyse sémantique du langage écrit, dont la famille dite BERT est la plus citée dans la littérature scientifique récente, atteignent des performances proches du niveau humain sur la classification d'événements industriels à partir de comptes rendus libres, sur corpus aviation (ASRS, CADORS), HAZOP, accidents de pipelines et accidents Seveso. La promesse commerciale converge : passer du suivi des indicateurs réactifs à une logique de *leading indicators* prédictifs, par corrélation transversale des sources internes.

Tensions FOH principales

Sur la vigilance. Le risque le plus documenté est l'installation d'une division tacite du travail de surveillance : « l'IA surveille les données, moi je surveille le terrain ». Cette répartition paraît raisonnable, mais elle suppose que chacun fait sa part. Or l'IA ne voit que ce pour quoi elle a été entraînée : un scénario inédit, une combinaison de facteurs jamais rencontrée resteront invisibles pour elle. Et la présence du capteur instrumental tend à éroder, insensiblement, l'attention humaine qui percevait le signal avant que la machine ne l'enregistre. La complémentarité réelle entre capteur humain et capteur machine suppose une conception délibérée — ne pas afficher le résultat instrumental avant que l'opérateur n'ait exprimé son propre ressenti, par exemple — et non une simple juxtaposition.

Sur la compréhension partagée. En situation d'urgence, l'équipe doit se construire rapidement un sens commun. Si l'IA produit une alerte issue d'une corrélation que personne ne sait expliquer, le sensemaking collectif est retardé ou faussé. Et si le modèle a dérivé sans que personne ne s'en aperçoive, les alertes deviennent peu fiables sans que le système le signale lui-même. L'expérience invite à outiller le système pour qu'il détecte et signale ses propres dégradations — et à préserver la capacité de l'équipe à fonctionner sans lui.

Sur le blanchiment de légitimité. Le risque ne se concentre pas ici sur une signature finale, comme en Q1, mais sur le statut du *score d'apprentissage* produit par l'IA. Un événement positionné par la machine en haut de l'axe « potentiel d'apprentissage » peut être validé comme tel par le Codir sans que personne ne sache reformuler en propre pourquoi il l'est. La proposition algorithmique tient alors lieu de jugement, et l'arbitrage qui aurait dû avoir lieu — entre les membres du Codir, sur ce que l'organisation considère effectivement comme apprenable — n'a pas eu lieu. On peut anticiper que la parade tienne moins au paramétrage de l'IA qu'au rituel managérial qui suit : interdire la ratification implicite, exiger une reformulation explicite du « pourquoi » par un membre du Codir avant tout classement définitif.

Sur la culture juste. L'analyse sémantique des comptes rendus rédigés par les équipes pose une question spécifique. Ces comptes rendus ne sont rédigés honnêtement que si leurs auteurs

ont la conviction qu'ils ne seront pas utilisés contre eux. Si l'exploitation algorithmique des verbatim devient un objet de surveillance individuelle — repérage des rédacteurs « négligents », classement des équipes par tonalité de leurs comptes rendus — la confiance qui en garantit la qualité s'érode, et la matière même que l'IA exploite se dégrade. C'est une boucle vicieuse à éviter par une doctrine claire : l'analyse des verbatim est traitée à un niveau collectif et anonymisé, jamais à des fins individuelles, et cette règle est inscrite dans une charte opposable.

Point d'attention : Le triage par la seule gravité immédiate est un filtre qui se présente comme neutre alors qu'il porte un choix doctrinal lourd : il prélève dans le flux ce qui ressemble à un accident *déjà identifié*, et laisse passer ce qui ne ressemble à rien de connu. Or les vrais signaux faibles sont précisément ceux qui ne ressemblent à rien de connu. La matrice à deux axes — gravité et potentiel d'apprentissage — n'est pas un raffinement technique, c'est une décision sur ce que l'organisation considère comme apprenable.

Où placer le curseur

- **Déléguer** — le filtrage du bruit, la déduplication des alarmes chroniques, le pré-classement des événements selon les scénarios connus.
- **Assister** — la corrélation multisource, l'analyse sémantique des verbatim, le positionnement sur la matrice à deux axes. L'IA propose des inférences ; l'analyste évalue, contextualise, arbitre. Niveau d'assistance *conseiller* en priorité ; le niveau d'assistance *producteur* est réservé aux fonctions de pré-classement sur des scénarios déjà documentés.
- **Sanctuariser** — la qualification finale d'un événement comme à fort potentiel de gravité ou d'apprentissage. C'est un acte managérial — souvent collectif — qui n'a de valeur que par l'échange contradictoire qu'il provoque. La signature en Codir d'un classement HIPO n'a pas vocation à être ratifiée sur la base du score : elle a vocation à être délibérée à partir du score.

Et pour l'entreprise étendue ? Le silence organisationnel des sous-traitants reste le filtre le plus puissant et le plus dangereux entre un signal faible et sa remontée. L'IA peut faciliter techniquement le signalement, mais elle ne résout en rien l'asymétrie contractuelle qui inhibe la parole. Un canal de remontée n'a de valeur que s'il est contractuellement et techniquement ouvert aux intervenants des entreprises extérieures dans les mêmes conditions de protection que pour les salariés internes, sans transit obligé par leur propre hiérarchie. À défaut, l'IA n'assemble que les signaux de la moitié de l'installation.

4.3 Q3 : Comment assister la conduite en temps réel sans se substituer au jugement de l'opérateur ?

La conduite en temps réel — salle de commande, intervention critique, gestion d'un transitoire imprévu — est le moment où la sécurité se joue en quelques secondes. Historiquement, l'opérateur affrontait une alarme brute et construisait son propre diagnostic à partir de ce qu'il voyait, de ce qu'il entendait, et de la mémoire de ses pairs. Avec les plateformes de données industrielles augmentées par IA — fusion temps réel SCADA/GMAO/vidéo, détection d'anomalies multivariée, jumeaux numériques opérationnels —, l'opérateur ne reçoit plus une information isolée, mais une situation déjà structurée et interprétée par une architecture complexe.

Le défi est cognitif autant que technique : l'IA façonne le modèle mental de l'opérateur avant qu'il n'ait pu se forger le sien. Si le système se trompe, hallucine, ou dérive silencieusement parce que la situation sort de son domaine d'apprentissage, l'opérateur sous pression temporelle manque du recul nécessaire pour contester une recommandation algorithmique formulée avec l'assurance d'une vérité mathématique.

Ce que l'IA peut apporter

Trois fonctions distinctes méritent d'être traitées séparément, parce qu'elles n'appellent ni les mêmes garanties ni les mêmes garde-fous.

Fonction 1 — Filtrer la saturation et organiser l'information. L'IA peut regrouper les alarmes liées, signaler les récurrences chroniques, vérifier la conformité formelle de séquences via capteurs d'état ou vision par ordinateur. Elle réduit massivement la surcharge cognitive en situation perturbée — « l'arbre de Noël » des alarmes redevient lisible. Cet apport est le plus mature et le plus immédiatement déployable. Mais le silence d'un système probabiliste qui n'a rien détecté n'est pas une garantie de normalité : c'est seulement le constat que rien dans son périmètre d'apprentissage n'a déclenché d'alerte. En industrie à risque, l'absence d'alerte algorithmique n'est jamais à elle seule une autorisation formelle de poursuivre.

Fonction 2 — Contextualiser l'anomalie par la plateforme transverse. L'IA croise en temps réel la dérive détectée avec l'historique de maintenance, les interventions en cours, le REX disponible, et offre une vue à 360° du contexte d'apparition de l'anomalie. Le rondier reçoit sur tablette une alerte contextualisée avant même d'arriver physiquement devant l'équipement ; l'ingénieur sûreté dispose d'un accès quasi instantané aux scénarios historiques comparables. L'apport est réel — à condition que l'intervenant de terrain n'en vienne pas à réduire son activité à la validation de ce qui s'affiche sur son écran. L'IA reste, par conception, aveugle à ce qu'aucun capteur ne mesure : une odeur inhabituelle, une vibration atypique, un micro-suintement, une coactivité hors planning. La vigilance sensible de l'intervenant est l'ultime capteur de ces signaux ; sa désactivation n'est jamais signalée par le système.

Fonction 3 — Aider au diagnostic en situation dégradée. Face à un transitoire complexe, l'IA — ou le jumeau numérique — peut proposer des hypothèses de défaillance et simuler instantanément les conséquences de différentes options de conduite. C'est l'apport le plus ambitieux et le plus délicat. Si le diagnostic proposé est opaque — « l'algorithme recommande de fermer la vanne X sans expliquer la chaîne causale » —, l'équipe peut exécuter sans comprendre. L'interprétation collective s'effondre, la décision est prise, mais l'intelligence de situation du collectif est détruite. Une recommandation algorithmique fonctionnant en boîte noire est, sur ce type d'usage, structurellement inacceptable.

Ce que cela suppose du modèle de sécurité

L'introduction de l'IA en conduite temps réel pose, plus qu'aux autres étapes, une question de frontière. La doctrine française de qualification des outils de calcul intervenant dans la démonstration de sûreté nucléaire[2] repose sur le caractère démontrable des outils. Cette doctrine n'a pas été conçue pour des modèles probabilistes dont le comportement varie avec les données qu'ils ingèrent. Les fonctions de protection automatique — arrêt d'urgence, mise en sécurité — restent du ressort des automatismes déterministes qualifiés ; l'IA probabiliste ne s'y substitue pas et ne s'y substituera pas tant que la doctrine de qualification n'aura pas évolué.

Dans cet espace contraint, l'IA peut intervenir sur des fonctions d'assistance — filtrage, contextualisation, aide au diagnostic — à condition que son périmètre fonctionnel soit explicitement borné, que sa traçabilité soit complète, et que son éventuelle dérive silencieuse soit elle-même surveillée. La détection de dérive du modèle — c'est-à-dire la surveillance continue de la qualité de ses prédictions au regard de la réalité observée — devient ici aussi importante que la fonction qu'elle protège : un modèle qui dégrade ses performances sans le signaler produit une assistance dont la fiabilité décroît sans que personne ne le sache.

Ce vers quoi tend le marché : Les grandes plateformes industrielles proposent aujourd'hui des briques opérationnelles — détection d'anomalies multivariée native (AWS IoT SiteWise, généralement disponible depuis juillet 2025), jumeaux numériques opérationnels (AWS IoT TwinMaker), assistants génératifs intégrés aux flux métier (Amazon Bedrock, Microsoft Fabric, Siemens Industrial Edge). Ces briques sont déployées dans la fabrication en série, l'énergie non critique et la logistique. Leur transposition aux industries à haut risque est annoncée par les fournisseurs, mais bute aujourd'hui sur deux obstacles : la qualité des données d'apprentissage — Deloitte [2025] rapporte que 70% des industriels identifient la donnée comme leur principal

[2] Doctrine formalisée par le Guide ASN n° 28 de 2017.

frein d'adoption — et l'exigence de démonstration de sûreté que la doctrine de qualification ne sait pas encore adresser pour des fonctions probabilistes. La pression à transposer ces capacités viendra. La question n'est pas de la refuser, mais de qualifier où s'arrête l'assistance et où commence ce qui ne peut pas relever d'un modèle probabiliste.

Tensions FOH principales

Sur la **compétence**. À force de superviser une IA qui gère brillamment les variations courantes, l'opérateur perd la dextérité de son raisonnement expert. Le jour où l'IA tombe en panne, dérive, ou affronte un événement hors de son apprentissage, l'humain doit reprendre la main à froid, en situation de stress intense, avec des compétences cognitives qui se sont atrophiées faute de pratique quotidienne. La parade tient à des exercices réguliers de conduite sans assistance — au simulateur pour les fonctions critiques, en exploitation courante pour le maintien des gestes — et à une conception de l'IA qui externalise son raisonnement (sources, niveau de confiance, limites connues du modèle) plutôt qu'elle ne le dissimule.

Sur la **compréhension partagée**. Quand le diagnostic est produit par une machine, l'équipe peut exécuter sans construire un sens commun. La parade est l'explicabilité — non pas comme exigence technique abstraite, mais comme condition opérationnelle de la décision : aucune recommandation algorithmique ne devrait piloter une situation accidentelle sans afficher la chaîne d'inférence qui l'a produite, les données qui l'alimentent, et les limites de validité du modèle qui la formule.

Sur le **blanchiment de légitimité**. Le risque ici ne se concentre pas sur une signature, mais sur l'asymétrie de responsabilité en situation de stress. Suivre l'avis algorithmique couvre ; le contester demande de justifier l'écart sous pression temporelle. La parade tient à une protection institutionnelle explicite du droit de désobéir — l'opérateur qui s'écarte d'une recommandation algorithmique au nom de son jugement expert ne devrait pas avoir à se justifier *a posteriori* du seul fait que la machine avait statistiquement raison. Sans cette protection, le droit de passe outre existe nominalement, mais devient inopérant en pratique.

Sur la **dérive silencieuse**. Un modèle probabiliste peut dégrader ses performances sans qu'aucun seuil d'alarme ne le signale. Cette dérive est d'autant plus dangereuse qu'elle est cohérente avec ce que le système rapporte — l'organisation croit voir ce que l'IA ne voit plus. La parade exige un dispositif de surveillance du modèle lui-même, distinct du système qu'il assiste.

Point d'attention : La frontière entre l'assistance acceptable et la substitution problématique ne tient pas à la sophistication du modèle, mais à la nature de la fonction exercée. Une IA très simple qui prend une décision de mise en sécurité franchit une frontière qu'une IA très sophistiquée qui éclaire un opérateur ne franchit pas. C'est la fonction qui qualifie le régime et le niveau d'assistance, pas la technologie. Cette distinction est aujourd'hui peu visible dans les discours commerciaux des fournisseurs ; elle reste pourtant l'opérateur principal de la doctrine de sûreté.

Où placer le curseur

- ▷ **Déléguer** — l'organisation et la hiérarchisation des informations, le filtrage des alarmes chroniques, la déduplication. L'IA y excelle, et la charge cognitive épargnée à l'opérateur est l'un des apports les plus tangibles.
- ▷ **Assister** — la contextualisation des anomalies et l'aide au diagnostic en situation dégradée. Niveau d'assistance *conseiller* exclusivement, sous condition stricte d'explicabilité. Le niveau d'assistance *producteur* est, pour ces fonctions, écarté : il reviendrait à confier à un modèle probabiliste un rôle que la doctrine de qualification réserve aux automatismes déterministes.
- ▷ **Sanctuariser** — les décisions de repli de l'installation, d'arrêt d'urgence ou d'évacuation. Ces fonctions restent hors algorithme et relèvent d'automatismes déterministes qualifiés ou de décisions humaines protégées. Le droit de l'opérateur à débrayer l'IA — droit de passer outre — doit être institutionnellement protégé et techniquement simple à exécuter.

Et pour l'entreprise étendue ? Dans de nombreuses installations, la salle de commande augmentée est opérée par le donneur d'ordre, tandis que l'exécution physique des manœuvres ou de la maintenance d'urgence est confiée à des sous-traitants. Si le centre de décision

impose une injonction de conduite issue de l'IA, mais que l'intervenant extérieur affronte une matérialité de terrain contradictoire sans avoir accès aux mêmes données ni au même droit de contestation — par peur de perdre le contrat —, l'IA crée une fracture de la représentation partagée. Le sous-traitant exécute un ordre algorithmique, la réalité physique du chantier n'a plus de voix. La symétrie d'accès aux données et de droit de contestation est ici une condition de la sécurité, non une concession contractuelle.

4.4 Q4 : Comment superviser un chantier à risque sans transformer la protection en surveillance ?

Sur un chantier à risque — zone contrôlée nucléaire, espace confiné, levage critique, intervention sur réseau sous pression —, le danger est souvent invisible : radiologique, chimique, anoxique, énergétique. Il ne se découvre historiquement qu'au franchissement d'un seuil de mesure ou d'une alarme dédiée. L'IA modifie cette donne en permettant d'analyser la dynamique combinée de plusieurs capteurs (dosimétrie, ambiance, position, vision) pour anticiper une dérive avant l'accident.

L'apport est réel : un modèle entraîné sur des séries temporelles ne se contente pas de comparer une valeur à un seuil, il reconnaît une *combinaison* de signaux qui, prise isolément, ne déclencherait rien. Une accélération de la dose conjuguée à une évolution de l'ambiance et à un repositionnement de l'intervenant peut signaler une situation que la lecture de chaque capteur séparément aurait laissée passer.

Mais c'est précisément dans cette supervision que l'extension fonctionnelle de la technologie devient la plus délicate à gouverner. Une caméra installée pour vérifier la séquence de consignation peut, par extension algorithmique, servir à surveiller le port des EPI, puis la posture, puis les temps de pause, jusqu'à l'identification nominative. Le dispositif matériel reste le même ; la finalité change. Ce glissement n'est pas un effet pervers occasionnel — c'est le mode de croissance économique normal d'une technologie de capture, qui maximise le retour sur investissement en multipliant les usages.

Ce que l'IA peut apporter

Trois fonctions méritent d'être traitées séparément, parce qu'elles n'engagent pas le même rapport à la personne supervisée.

Fonction 1 — Détecter précocement les dérives. Au lieu d'attendre une alarme de seuil, un modèle entraîné sur des séries temporelles reconnaît une dynamique combinée et propose un repositionnement anticipé. L'intervenant bénéficie d'un filet de sécurité proactif. Le revers tient à un déplacement insensible de sa propre vigilance : habitué à se fier à des capteurs communicants, il peut perdre la compétence de lire son environnement de manière experte — propreté, marquage, signaux d'ambiance non instrumentés. La vigilance instrumentée et la vigilance personnelle ne se cumulent pas spontanément ; leur articulation suppose une conception délibérée du dispositif.

Fonction 2 — Vérifier des gestes critiques. La vision par ordinateur peut confirmer qu'une zone d'exclusion est respectée, qu'un EPI critique est porté… Elle ajoute une barrière au moment le plus vulnérable de l'intervention. C'est, sur des fonctions précisément bornées, l'apport le plus défendable de la vision par ordinateur — y compris par ceux qui sont par ailleurs critiques du déploiement de caméras sur les chantiers. La condition de cette légitimité tient à la définition du périmètre fonctionnel : la machine certifie un geste de sécurité défini en amont, elle n'évalue pas un comportement.

Fonction 3 — Piloter à distance. Le superviseur hors zone dispose d'une vue synthétique — répartition spatiale, évolution des ambiances, avancement du scénario. Il voit le contexte global que l'intervenant, concentré sur sa tâche, ne perçoit pas. Le revers est connu : en pilotant exclusivement depuis son tableau de bord, le responsable perd le contact avec le travail réel, la charge physique, l'imprévu. Il peut formuler des injonctions déconnectées de la matérialité du terrain. Le pilotage à distance augmente la vision panoramique, il ne remplace pas la présence.

Exemple

Temps-critique et réduction du risque résiduel : quand l'IA commande l'évacuation

Sur chantier, la fenêtre temporelle entre la détection d'un danger et le moment où il se matérialise peut-être inférieure au temps nécessaire pour qu'un opérateur perçoive le signal, décide et agisse. Dans ces situations, on peut anticiper que confier l'ordre d'évacuation à un système de détection et de décision algorithmique réduit significativement le risque d'exposition prolongée — et que ce gain en sécurité substantielle prime sur les considérations d'organisation de la décision.

Un chantier en zone explosible (atmosphère explosive, poussière, gaz) où un capteur détecte un confinement involontaire ou une dérive de paramètre vers la zone dangereuse. Le temps disponible avant formation d'une atmosphère explosible est souvent de l'ordre de quelques secondes. Un opérateur, même attentif, met 200 à 300 ms rien que pour percevoir l'alerte, puis décider d'ordonner l'évacuation. Pendant ce délai, le danger progresse. À l'inverse, un système de détection-décision agit en 50-100 ms. Cette différence — un ordre de grandeur — peut faire la différence entre une exposition contrôlée et un accident.

Ce que cela suppose du modèle de sécurité

La supervision augmentée par IA rend opérationnelle une question que les organisations préféraient souvent laisser implicite : quelle est la finalité légitime des données collectées sur le travail ? Tant que la dosimétrie servait exclusivement à protéger l'intervenant des rayonnements ionisants, la finalité était évidente et la confiance pouvait s'installer sans contrat explicite. Avec des dispositifs qui mesurent simultanément la dose, la position, la posture, le geste et la durée, la finalité ne va plus de soi — elle doit être explicitée, bornée, et opposable.

Cette explicitation suppose trois choses. D'abord, une définition publique du périmètre fonctionnel : ce que l'algorithme est autorisé à détecter, ce qu'il ne l'est pas, et qui valide les évolutions de ce périmètre. Ensuite, une asymétrie d'accès assumée : l'intervenant supervisé doit pouvoir consulter les données qui le concernent, sans pour autant que le superviseur dispose d'un accès individuel à des fins évaluatives. Enfin, une instance de contrôle externe : la décision d'étendre les capacités du modèle à d'autres fins que sa finalité initiale ne peut pas relever de la seule maîtrise d'ouvrage, parce que c'est précisément le point où les intérêts économiques et les intérêts de sécurité divergent.

Le RGPD, la directive-cadre européenne sur la santé et la sécurité au travail, et les exigences renforcées de l'AI Act sur les systèmes à haut risque convergent sur cette ligne. Aucun de ces textes ne suffit isolément à gouverner le glissement ; leur articulation locale, par installation et par chantier, reste à construire.

Ce vers quoi tend le marché : Les briques techniques de la supervision augmentée sont disponibles et mûrissent rapidement : vision par ordinateur entraînée sur des gestes industriels, capteurs IIoT portés à dosimétrie connectée, plateformes de pilotage temps réel. Le marché propose ces briques avec des cas d'usage de plus en plus larges — sécurité, conformité, productivité, optimisation — sans toujours distinguer les régimes de finalité qui les séparent. La distinction est pourtant la ligne où se joue l'acceptabilité sociale du dispositif : un même flux vidéo peut servir une vérification ciblée ou une surveillance généralisée. La technologie n'arbitre pas — c'est l'organisation qui décide, et qui assume.

Tensions FOH principales

Sur la **compétence**. L'intervenant qui s'habitue à des capteurs communicants peut voir s'éroder la compétence de lecture experte de son environnement — propreté, marquage, signaux d'ambiance, comportement collectif. Cette érosion est d'autant plus délicate à compenser qu'elle est progressive et qu'elle n'apparaît qu'au moment où le capteur défaille. La parade tient à des moments de travail délibérément non instrumentés — tournées de site, exercices de lecture sensible — qui ne sont pas une régression, mais une conservation active de la compétence.

Sur le **blanchiment de légitimité**. Quand la machine certifie un geste, la validation humaine du même geste devient redondante en apparence. Elle ne l'est pas. Le blanchiment opère ici par dilution : si la responsabilité formelle reste portée par un opérateur, mais que la vérification effective est entièrement déléguée à l'algorithme, l'opérateur signe en réalité une décision qu'il n'a pas instruite. La parade : une articulation explicite entre vérification algorithmique et validation humaine, avec une plus-value distincte de la seconde — typiquement, la vérification de ce que l'algorithme n'a pas pu voir, plutôt que la confirmation de ce qu'il a vu.

Sur la **culture juste**. C'est la tension la plus sensible de cette question. L'introduction de la vision par ordinateur sur un chantier crée par construction une asymétrie de regard : la machine voit l'intervenant, l'inverse n'est pas vrai. Si le dispositif glisse — même sans intention initiale — vers la détection comportementale ou l'évaluation de productivité, l'opérateur se sent tracé et scoré. La confiance qui rendait la déclaration sincère possible se dégrade, et le système d'apprentissage qui dépendait de cette déclaration s'appauvrit. La parade ne tient pas à des engagements moraux, mais à des dispositifs vérifiables : périmètre fonctionnel contractualisé avec les instances représentatives du personnel, audit indépendant des usages effectifs, droit de retrait du dispositif sans préjudice contractuel pour les sous-traitants concernés.

Sur la **compréhension partagée**. Quand l'intervenant reçoit une injonction d'évacuation issue d'un modèle probabiliste qu'il ne comprend pas, il peut obéir sans construire le sens de la situation. La parade rejoint celle de Q3 : explicabilité opérationnelle de la recommandation — sources, niveau de confiance, limites du modèle — et droit institutionnellement protégé de contester l'algorithme sans avoir à se justifier *a posteriori* si la machine avait statistiquement raison.

Point d'attention : Un même flux vidéo peut nourrir une vérification ciblée ou une surveillance comportementale généralisée. Ce qui en détermine la nature n'est pas la technologie : c'est la décision explicite sur la finalité, le périmètre fonctionnel, et la gouvernance de la lecture des images. Le dispositif technique n'est pas doctrinalement neutre — il porte les usages dans lesquels il a été conçu. Mais le dispositif organisationnel qui l'encadre est, lui, celui qui décide à chaque instant lesquels de ces usages sont mis en œuvre.

Où placer le curseur

- **Déléguer** — la centralisation des mesures issues des capteurs, la déduplication des alertes, le prépositionnement des hypothèses de dérive sur des configurations connues.
- **Assister** — la vérification ciblée des gestes critiques (niveau d'assistance *conseiller*, périmètre fonctionnel contractualisé) et le pilotage à distance (niveau d'assistance *conseiller* exclusivement, en complément de la présence terrain et non en substitution). Le niveau d'assistance *producteur* — décision automatique sur la base de la sortie algorithmique — est écarté pour ces fonctions, parce que la matérialité du chantier ne se laisse pas valider à distance.
- **Sanctuariser** — trois objets distincts :
  1. **Les consignes du superviseur quand le temps de réponse le permet** : L'appréciation d'une situation de danger immédiat et l'ordre d'évacuation relèvent du superviseur direct ou de l'intervenant, qui reçoit de l'algorithme une recommandation structurée (alerte, contexte, options). Le droit d'écart reste simultané et institutionnellement protégé : le superviseur peut contester, ajuster ou refuser l'ordre algorithmique sans justification préalable ni surcoût procédural.
  2. **Quand le temps de réponse est trop court : la délégation au système** : Dans les configurations où le temps minimal requis pour percevoir, décider et agir est inférieur à la fenêtre temporelle disponible avant matérialisation du danger, on peut anticiper que l'ordre d'évacuation sera émis directement par le système algorithmique, sans intervention humaine préalable. Cette délégation suppose : (a) une qualification préalable et contractualisée des seuils de déclenchement, (b) une traçabilité complète de chaque décision et de sa justification, (c) un droit de contestation *ex post* de la pertinence de l'ordre, exercé par un acteur tiers et remontant annuellement au comité de sécurité et au dialogue social.
  3. **Le périmètre fonctionnel de l'algorithme lui-même** : son extension à de nouvelles finalités ne peut pas relever d'une décision technique, elle exige une nouvelle décision doctrinale, contractualisée avec les instances représentatives du personnel.

La supervision instrumentée (humaine, algorithmique ou hybride) a pour finalité première la détection en temps réel d'écarts pendant l'intervention. Mais elle produit, comme effet secondaire, une amélioration de la capture des situations à haut potentiel de gravité — celles qui auraient pu produire un accident grave sans en produire. Cette contribution à la matière du REX, sans être l'objectif visé, est un acquis à reconnaître — il sera mobilisé en Q5, où l'on verra que la limite de la capture cesse d'être bloquante dès lors que l'analyse approfondie des événements collectés va jusqu'aux causes systémiques.

Et pour l'entreprise étendue ? Sur les chantiers opérés par des salariés d'entreprises extérieures, la supervision en temps réel crée un risque d'ingérence : le donneur d'ordre devient de fait superviseur opérationnel, ce qui altère la relation d'emploi légale.

Un principe à sanctuariser : les salariés des entreprises extérieures sont supervisés par un salarié de leur propre employeur ou par un tiers de confiance désigné contractuellement. Le donneur d'ordre ne reçoit qu'une information consolidée, sauf pour les alertes jugées vitales (gaz, explosion, confinement) qui sont partagées directement.

Le tiers de confiance assume deux responsabilités : transmettre les alertes tactiques aux superviseurs légaux, et assurer que ces alertes ne deviennent pas des outils de contrôle du donneur d'ordre sur les sous-traitants. C'est une clarification de responsabilités, pas une protection technique contre le glissement vers la surveillance.

Le compromis : le donneur d'ordre accepte une limitation sur sa visibilité temps-réel — c'est le prix pour maintenir le lien d'emploi légal et la confiance opérationnelle dans l'équipe étendue.

4.5 Q5 : Comment transformer les analyses en apprentissage organisationnel ?

L'industrie à haut risque dispose de systèmes de REX très structurés. Mais le constat reste sévère : le REX ne se referme pas toujours sur une transformation des barrières. Les organisations accumulent les analyses d'événements individuels, corrigent les écarts immédiats, et peinent à croiser ces analyses pour questionner les hypothèses de leur modèle de sécurité.

La première raison de cet écart n'est pas informationnelle. Elle tient à la posture du commanditaire — directeur de site, directeur métier, membre du Codir. Quand le commanditaire attend un rapport à valider plutôt qu'une exploration à conduire, l'analyse se referme prématurément, contourne les zones d'inconfort (FOH, arbitrages production/sécurité, règles inapplicables) et débouche sur un plan d'actions qui rassure la gouvernance sans interroger le modèle. À l'inverse, un commanditaire curieux ouvre l'espace d'analyse, protège la culture juste et transforme un rapport en évolution tangible du système. **La curiosité du commanditaire fait la profondeur de l'analyse** — c'est un acte managérial délibéré, antérieur à tout outil. Aucune technologie ne crée cette curiosité ; et si elle n'est pas là, **l'IA n'a aucun moyen de la suppléer**.

La distinction de **[Argyris et Schön 1978]** entre *apprentissage en simple boucle* et *apprentissage en double boucle* éclaire ce que l'on attend du REX. Dit simplement : la simple boucle traite l'accident — on corrige l'écart, on remet la barrière en place, on rappelle la règle ; la double boucle traite le type d'accident — on remonte aux causes organisationnelles et systémiques qui rendront d'autres événements possibles tant qu'elles ne seront pas instruites. L'expérience montre que les analyses unitaires vont rarement jusqu'à la double boucle, même quand elles s'en donnent la forme.

Sur ce socle, trois difficultés concrètes freinent la transformation des analyses en apprentissage :

- **Collecter et choisir**. Les organisations à haut risque ne captent pas toutes leurs situations à haut potentiel de gravité. La littérature situe autour de 20% la part des HIPO remontés par la seule déclaration spontanée des salariés dans les meilleures organisations **[Amalberti 2013]**. La supervision — humaine ou algorithmique (voir encadré Q4) — élève ce taux. Mais ce n'est pas le point décisif. Le point décisif, qu'il faut entendre tel qu'Amalberti le formule, est que les 20% captés *suffisent* à apprendre — à condition que l'analyse soit conduite jusqu'aux causes systémiques, communes aux 80% invisibles. Reste alors à choisir ce qui, parmi le collecté, mérite l'analyse approfondie. Sans critère explicite de tri — gravité réelle ou potentielle croisée avec potentiel d'apprentissage — l'organisation approfondit *en moyenne*, c'est-à-dire jamais vraiment.
- **Approfondir vraiment**. Pour les événements qui justifient la double boucle, l'analyse doit couvrir les axes pertinents (technique, humain, organisationnel, culturel) et remonter au-delà des conditions locales (N1-N2) jusqu'aux mécanismes organisationnels (N3) et, quand l'événement le justifie, aux choix de gouvernance (N4). C'est rarement le cas : le facteur organisationnel est systématiquement sous-instruit, le facteur culturel pratiquement absent, la séparation faits/hypothèses/inconnues n'est pas tenue, les barrières de récupération sont le point aveugle récurrent. Le résultat est un format de synthèse qui porte l'étiquette

d'analyse approfondie sans en remplir les exigences — et qui, par sa diffusion régulière, finit par fixer la norme.

- ▷ **Croiser pour interroger le modèle**. L'analyse unitaire, même bien conduite, ne suffit pas. Ce qui désigne une vulnérabilité systémique, ce sont les régularités qui n'apparaissent qu'en croisant plusieurs analyses : interfaces récurrentes, barrières systématiquement faibles sur une famille d'activités, scénarios qui se ressemblent à travers des événements en apparence sans lien. Ce niveau de relecture — *analyse de niveau 2* — exige une matière de qualité, du temps, et un format qui n'est pas celui de l'analyse cas par cas.

La boucle courte transforme un signal en décision visible au poste de travail ; la boucle longue recalibre le modèle de sécurité au niveau de la gouvernance. C'est dans ce double système — quoi questionner, à quelle cadence — que l'IA peut intervenir, avec des régimes (DAS) et niveau d'assistance (outil, conseiller, producteur) différents selon le niveau et le moment.

Ce que l'IA peut apporter

Les trois difficultés appellent trois familles d'apports — capter et coter, outiller l'analyse approfondie, croiser —, auxquelles s'ajoute un effet transversal que l'auteur tient pour central : la **fonction miroir**, par laquelle l'IA révèle le modèle de sécurité que l'organisation mobilise pour analyser ses événements.

Fonction 1 — Capter et coter. L'IA peut explorer les bases de remontées (situations dangereuses, comptes rendus libres, mains courantes) pour proposer à un commanditaire humain des candidats à devenir des HIPO. Le traitement du langage naturel rend exploitables des champs textuels invisibles à la lecture séquentielle, et la littérature confirme la maturité de cette fonction pour la classification d'événements selon la gravité et la typologie **[Wang et Eisner 2017]**. Sur les candidats identifiés, l'IA peut produire une cotation préliminaire — gravité réelle ou potentielle, potentiel d'apprentissage — et proposer une hiérarchisation **[Quiot 2026]**. Limite à tenir : l'IA n'identifie que ce qui a été écrit. Si le silence organisationnel filtre déjà ce qui remonte, la fonction *capter et coter* augmente la capture dans la matière disponible sans toucher à ce que la culture juste, en amont, autorise à mettre par écrit.

Fonction 2 — Outiller l'analyse approfondie. C'est ici que l'apport de l'IA est le plus mature et le plus immédiatement utile. Trois usages se cumulent. *D'abord*, consolidation factuelle : croiser automatiquement les données de supervision (SCADA, alarmes), l'historique de maintenance et les comptes rendus libres pour livrer à l'analyste une matière contextualisée. *Ensuite*, premier jet : sur les événements récurrents standardisés, l'IA peut proposer un dossier préconstruit avec des causes apparentes candidates, que l'analyste révise plutôt qu'écrit — bascule qui n'est pas neutre et qu'on traite en *Tensions*. *Surtout*, et c'est l'apport le plus original, l'IA peut tenir un **rôle socratique** : interroger l'exhaustivité des axes couverts, signaler les sauts de causalité, vérifier la séparation faits/hypothèses/inconnues, tester la profondeur atteinte (N1-N4), proposer des questions complémentaires pour éviter l'explication spontanée. Ce déplacement est doctrinalement important : l'IA n'est pas là pour *trouver* les causes, elle est là pour *empêcher l'analyste de s'arrêter trop tôt*. La compétence d'analyse reste humaine ; l'IA en outille la rigueur méthodologique.

Fonction 3 — Croiser et faire émerger des patterns. L'IA peut ingérer des centaines d'analyses pour faire émerger des configurations récurrentes qu'aucune lecture cas par cas ne peut produire — interfaces fragiles, barrières systématiquement faibles, conjonctions précises entre météo, coactivité et affaiblissement d'une défense. Elle peut aussi identifier le *REX positif* : les adaptations réussies par le terrain face à l'aléa, qui ne remontent pas parce qu'elles ne déclenchent pas d'événement à déclarer.

Mais l'IA atteint ici une limite structurelle : elle peut montrer que des facteurs convergent statistiquement — le *quoi* —, elle ne peut pas raisonner causalement — le *pourquoi*. **[Pearl 2009]** a formalisé cette frontière. Conséquence claire : ne pas attendre de la fonction *croiser* qu'elle découvre des causes profondes que les analyses individuelles n'auraient pas vues. Si les analyses unitaires sont superficielles, les patterns extraits seront superficiels — agrégés, pas plus profonds.

Fonction 4 — La fonction miroir, effet transversal. À chacune des trois étapes précédentes, l'IA produit, sans le chercher, un effet que l'auteur tient pour central. Quand elle cote des candidats HIPO, elle révèle ce que l'organisation considère comme grave et apprenant — son

modèle de risque. Quand elle questionne socratiquement, elle révèle les axes que l'organisation explore et ceux qu'elle évite — son modèle de causalité. Quand elle croise des analyses, elle révèle les régularités de pensée — le modèle de sécurité implicite à travers lequel toutes les analyses ont été produites. Cette fonction miroir déplace la question utile : ce n'est pas *« quelles causes profondes l'IA révèle-t-elle ? »*, mais *« quel modèle de sécurité l'IA donne-t-elle à voir, et l'organisation accepte-t-elle de le regarder ? »*. Elle borne aussi ce qu'il est raisonnable d'attendre : l'IA ne découvre pas les causes que les analyses humaines n'ont pas vues ; elle rend visible la manière dont l'organisation pense, formule et traite ses événements — et c'est déjà beaucoup, à condition que la délibération s'en empare.

Ce que cela suppose du modèle de sécurité

Deux préalables doivent être tenus avant déploiement :

- L'**explicitation du référentiel d'analyse** : pour que la cotation gravité × potentiel d'apprentissage produise des candidats HIPO pertinents, pour que le questionnement socratique teste une profondeur définie, pour que les patterns soient lisibles, l'organisation doit avoir formulé ce qu'elle considère comme une analyse suffisamment profonde — axes à couvrir, niveau N à atteindre, critères distinguant cause et symptôme, indicateurs d'analyse blâmante ou superficielle. Sans ce référentiel, le standard implicite mobilisé par l'IA est celui de ses concepteurs — qui n'est pas nécessairement celui de l'organisation, et qui de surcroît est invisible. Le référentiel ne préexiste pas dans la plupart des organisations ; sa construction est en elle-même un acte d'apprentissage organisationnel, préalable à l'IA et souvent plus structurant qu'elle.
- La **matière narrative** que l'IA va exploiter. Comptes rendus libres, débriefings, mains courantes portent les biais de leur production : ce qui est confortable à écrire y est surreprésenté, ce qui mettrait le rédacteur en difficulté y est sous-représenté. Une IA qui ingère ces données reproduit ces biais — et peut les amplifier si son usage devient lui-même un déterminant de la rédaction. La culture juste n'est pas seulement souhaitable : elle est opérationnellement préalable à toute exploitation à grande échelle de la matière narrative. Si elle n'est pas tenue, l'IA n'amplifie pas l'apprentissage, elle amplifie le silence.

Ce vers quoi tend le marché : Les plateformes d'analyse de REX et de safety analytics se multiplient — solutions natives des grands fournisseurs cloud (AWS, Microsoft, Google), spécialistes de la sécurité industrielle (Sphera, Intelex, Enablon), startups dédiées à l'analyse sémantique des verbatim. Toutes proposent peu ou prou la même promesse : corrélation transversale, identification de patterns, génération assistée d'arbres des causes, tableaux de bord exécutifs. Le marché propose les briques ; il ne propose pas — et ne peut pas proposer — le modèle de sécurité qui leur donne sens. C'est cette articulation qui reste à construire localement, chantier par chantier, organisation par organisation.

Tensions FOH principales

Sur la **compétence**. L'analyste qui passe du niveau d'assistance *conseiller* au niveau d'assistance *producteur* peut perdre la compétence de construire une analyse depuis la page blanche. Le risque est d'autant plus délicat qu'il se manifeste sur un événement atypique, où l'IA n'a pas de pattern à proposer. La parade : maintien périodique d'analyses non assistées, sur des événements de complexité moyenne — pas seulement sur les cas exceptionnels.

Sur le **blanchiment de légitimité**. Quand l'IA produit un score d'apprentissage, un classement HIPO ou une cotation, ce score peut tenir lieu de jugement. Le Codir ratifie ; il ne délibère pas. La décision de retenir un événement pour analyse approfondie, de modifier une barrière ou de réviser un équilibre du modèle est un jugement de valeur exclusif au Codir. La proposition algorithmique nourrit la délibération ; elle ne la remplace pas.

Sur la **culture juste**. Le croisement de données RH (ancienneté, formation, parcours) avec des données d'événements est un champ de mines. Si ce rapprochement est perçu comme un outil de mise en cause individuelle, la matière narrative se tarit. L'usage du REX outillé doit être doctrinalement borné à un niveau collectif et anonymisé, inscrit dans une charte opposable et accessible aux instances représentatives du personnel.

Sur la **propagation des biais d'analyse**. L'IA n'apprend des analyses que ce qu'elles donnent à voir. Si le corpus existant est superficiel — N1-N2 dominants, organisationnel sous-instruit,

formulations blâmantes prises pour causes — l'IA en généralise la superficialité avec une apparence d'objectivité algorithmique qui rend la correction plus difficile. La parade : qualifier la matière avant exploitation (filtrage de qualité, écart au référentiel d'analyse) ; instruire l'IA avec un thésaurus multi-modèles plutôt qu'aligné sur le seul modèle implicite de l'organisation.

Sur le **tarissement**. Si la rédaction des comptes rendus s'aligne progressivement sur ce que la machine traite bien, le corpus s'appauvrit. La parade tient à la préservation d'espaces de production narrative non outillés — débriefings, entretiens, récits libres — dont la valeur ne se mesure pas à leur exploitabilité algorithmique.

Point d'attention : Une IA qui détecte des patterns dans le REX produit du matériau, pas du sens. Le sens se construit dans la délibération qui suit, et c'est cette délibération qui constitue l'apprentissage organisationnel. Une organisation qui ratifie les sorties algorithmiques sans les délibérer n'apprend pas davantage qu'une organisation qui ignore son REX — elle se donne seulement l'apparence d'apprendre, ce qui est plus difficile à corriger qu'une simple absence d'analyse.

Où placer le curseur

- **Déléguer** — l'enrichissement contextuel du dossier d'événement (consolidation des sources, croisement automatique des données de supervision avec l'historique de maintenance), la déduplication, le pré-classement sur typologies connues. La capture de candidats HIPO (Fonction 1) peut être déléguée sur la *détection* (extraire de la masse les remontées qui méritent un regard humain) ; la *qualification* d'un événement comme HIPO reste une décision humaine.
- **Assister** — trois usages avec des positions distinctes sur le continuum *outil/conseiller/producteur*.
  - **Cotation** (Fonction 1) : Niveau d'assistance conseiller — l'IA propose, un commanditaire humain arbitre.
  - **Analyse approfondie** (Fonction 2) : niveau d'assistance *conseiller* sur les événements atypiques où le rôle socratique est l'apport décisif ; niveau d'assistance *producteur encadré* sur les événements récurrents standardisés. Le niveau d'assistance *producteur* exige un référentiel d'analyse explicité, le maintien périodique d'analyses non assistées, et une qualification préalable de la matière.
  - **Croisement** (Fonction 3) : niveau d'assistance *conseiller* exclusivement, parce que le passage du *quoi* au *pourquoi* n'est pas délégable et que la propagation des biais menace tout régime et niveau d'assistance plus engageant.
- **Sanctuariser** — l'arbitrage stratégique sur le modèle de sécurité. La décision de retenir un événement pour analyse approfondie, de modifier une barrière, de réviser un équilibre du modèle ou de qualifier l'acceptabilité d'un mécanisme révélé par l'IA est un jugement de valeur exclusif au Codir : à délibérer *à partir* du score, pas à ratifier *sur la base* du score. L'interprétation de la lecture-miroir (Fonction 4) est elle aussi sanctuarisée : ce que l'IA révèle du modèle de sécurité ne se confond jamais avec ce que le modèle *doit être* — cet arbitrage reste un acte managérial délibéré, lié à la curiosité du commanditaire posée en amont.

Et pour l'entreprise étendue ? L'apprentissage organisationnel à l'échelle d'un site industriel est structurellement biaisé si l'outil d'IA n'ingère que les données internes de l'exploitant. Le silence organisationnel des sous-traitants — souvent contraints par des enjeux de conservation de marché — prive la machine des signaux d'interface et des causes profondes liées à l'asymétrie DO/ST. La sous-traitance représente, sur les sites à risque, une part significative des heures travaillées ; un REX qui en exclut la matière apprend un système amputé. La réciprocité est une exigence pratique autant qu'éthique : un outil qui exploite les données du sous-traitant sans lui en restituer les enseignements transversaux institue une asymétrie d'apprentissage qui dégrade, sur la durée, la qualité même de la matière partagée.

4.6 Q6 : Comment piloter la robustesse plutôt que l'activité ?

Le pilotage de la sécurité s'appuie historiquement sur des indicateurs qui mesurent l'activité réalisée : volumes de maintenance, taux de conformité documentaire, taux de fréquence des accidents. Ces indicateurs ont une qualité — ils existent, ils sont consolidables, ils permettent la comparaison entre sites et entre périodes. Ils ont aussi une faiblesse : ils mesurent ce qui a été *fait*, pas ce qui a été *tenu*. Une organisation peut afficher d'excellents résultats sur ses indicateurs d'activité tout en voyant la santé réelle de ses barrières de défense se dégrader silencieusement.

Une distinction doctrinale plus profonde se cache derrière cette faiblesse. La prévention des accidents graves et mortels (PAGEM) ne se pilote pas avec les mêmes leviers que la baisse du taux de fréquence. Les indicateurs de sinistralité courante mesurent un résultat passé, sensible aux aléas individuels, et sont influencés à court terme par des leviers comportementaux. La PAGEM, elle, dépend de leviers systémiques qui agissent sur des cycles longs : robustesse des barrières critiques sur les scénarios à haut potentiel, qualité de la préparation des activités critiques, robustesse des interfaces avec les prestataires, traitement effectif des précurseurs. Passé un certain seuil de maturité, pousser encore sur la sinistralité courante peut même éroder la prévention du grave : la rigidification qui s'ensuit — multiplication des règles, traque des écarts de surface — érode la capacité d'adaptation, qui est précisément la ressource qui rend les barrières résilientes face à l'imprévu **[Amalberti 1996]**. Une organisation peut donc afficher des indicateurs favorables et rester très exposée à ses scénarios graves.

Pour piloter la prévention des accidents graves, il faut donc d'autres indicateurs que ceux de la sinistralité courante. Des indicateurs qui rendent visible la *robustesse des barrières issues du modèle de sécurité* — disponibilité conjointe des barrières critiques, état des règles qui sauvent, qualité effective du STOP chantier, densité des précurseurs sur les scénarios à haut potentiel, cohérence entre prescrit et opérant sur des familles d'activités à enjeu. Cette robustesse n'est pas directement observable ; elle s'infère d'un croisement de données qui dorment dans des systèmes séparés (GMAO, SCADA, SIRH, SI HSE) que personne ne consolide. C'est ici que l'IA peut apporter quelque chose — à condition de ne pas se contenter d'accélérer le *reporting* de conformité existant.

Point d'attention : La question des indicateurs n'est pas celle de leur contractualisation. Un indicateur de sinistralité courante éclaire utilement quand il est partagé, mais peut inverser son effet quand il est contractualisé sur un périmètre qui ne le maîtrise pas — c'est la loi de Goodhart. Cette question, qui dépasse le strict cadre de l'IA, est tenue distincte dans la suite de ce chapitre.

Ce que l'IA peut apporter

L'apport de l'IA en pilotage suit une progression naturelle, du plus consensuel au plus discutable. Cette progression mérite d'être tenue distincte, parce qu'elle correspond à trois régimes de gouvernance différents.

Fonction 1 — Produire des indicateurs descriptifs. L'IA peut décloisonner les bases de données — GMAO, SCADA, SIRH, HSE — pour construire des métriques de santé du système : délai de fermeture effective des actions REX, évolution de la couverture des habilitations croisée au planning, disponibilité conjointe de barrières critiques, écart entre les durées planifiées et les durées effectives des interventions sur les équipements à enjeu. Le manager de proximité gagne en pertinence : il ne vérifie plus que son équipe a coché les cases, il visualise quelles barrières tiennent et lesquelles s'affaiblissent. C'est l'apport le plus défendable, à condition que l'indicateur reste descriptif — il rend visible une situation, il ne prescrit pas l'action à mener.

Fonction 2 — Rendre lisible l'état de la robustesse. Les indicateurs descriptifs de F1 ne s'interprètent pas isolément. C'est leur imbrication qui dit quelque chose — la disponibilité d'une barrière critique lue en regard de la densité des précurseurs sur le même périmètre, croisée avec la qualité effective du STOP et le rythme des chantiers de coactivité. Ce n'est plus un indicateur qu'on lit, c'est un paysage qu'on parcourt.

L'IA peut ici outiller le manager d'un usage qui n'existait pas avant elle : le dialogue d'interrogation. Plutôt que de lire passivement un tableau de bord, le manager interroge en langage naturel les éléments sous-jacents — quelle barrière s'affaiblit ? sur quel périmètre ? depuis quand ? quels précurseurs ont monté en parallèle ? —, fait remonter la chaîne d'agrégation,

conteste une corrélation qui lui paraît douteuse, croise avec ce qu'il sait du terrain. Ce dialogue d'explicabilité, qu'on rencontre dans la littérature récente sous le nom de conversational XAI, rend l'indicateur composite contestable — propriété décisive pour un pilotage de la robustesse en environnement complexe.

Cet apport ne doit pas être confondu avec l'aide à la décision que de nombreux fournisseurs proposent aujourd'hui. Rendre lisible n'est pas recommander. La distinction n'est pas linguistique, elle est structurelle. La décision en sécurité industrielle mobilise simultanément quatre dimensions — la technique, le système de management, les facteurs humains et organisationnels, et la dimension culturelle. L'IA peut traiter la première avec une bonne maturité, la deuxième avec des limites, et bute sur les deux dernières : FOH et culture ne sont pas des données structurées qu'on encode, ce sont des phénomènes émergents qui se constatent dans le travail réel et qui ne sont accessibles qu'à un acteur engagé dans ce travail. La sécurité étant une propriété émergente du système sociotechnique et non une somme de paramètres mesurables **[Leveson 2004]**, une IA qui recommande une action à partir d'indicateurs — aussi sophistiqués soient-ils — opère sur la moitié du sujet et ignore l'autre. La parade n'est pas d'améliorer l'explicabilité de la recommandation ; elle est de ne pas la produire. L'IA rend lisible l'état de la robustesse ; le manager décide à partir de cette lecture et de ce que les indicateurs ne montrent pas.

Ce que cela suppose du modèle de sécurité

Un indicateur descriptif encode une thèse sur ce qui compte. Le choix des variables consolidées, leur pondération, leur agrégation, leur seuil d'alerte sont autant de décisions doctrinales déguisées en paramétrages techniques. L'organisation qui n'a pas explicité ce qu'elle considère comme une barrière critique, ou comme un précurseur significatif, se retrouve à le découvrir incarné dans les paramètres de son tableau de bord — sans l'avoir délibéré.

Une exigence émergente mérite d'être posée : l'audit périodique du modèle lui-même. Un indicateur composite construit pour rendre lisible l'état de la robustesse encode une représentation de ce qu'est la robustesse à un instant donné — quelles barrières sont critiques, quels précurseurs comptent, quels seuils signalent un affaiblissement. Si les conditions du système évoluent — changement de procédé, rotation du personnel, modification de l'organisation du travail, évolution de la sous-traitance — sans que l'indicateur soit recalibré, l'organisation lit une cartographie obsolète en croyant la lire à jour. Cette dérive est invisible par construction : elle ne se manifeste qu'au moment où l'écart entre ce que l'indicateur affiche et ce que le terrain montre devient flagrant, c'est-à-dire trop tard. Cette exigence d'audit n'a pas d'équivalent dans la doctrine classique du pilotage sécurité, qui supposait des indicateurs construits une fois pour toutes et stables dans le temps. Avec des modèles qui dérivent à mesure que les données qu'ils ingèrent évoluent, la doctrine doit intégrer la maintenance du modèle comme une fonction de sécurité à part entière.

Ce vers quoi tend le marché : Les plateformes de *safety intelligence* et de *risk analytics* — solutions natives des grands fournisseurs cloud (AWS, Microsoft, Google) comme spécialistes établis du HSE software (Sphera, Intelex, Enablon) — proposent aujourd'hui un éventail d'usages qu'il est utile de distinguer pour évaluer une offre. *Premier étage* : la consolidation transverse des données HSE, opérationnelles et RH pour produire des indicateurs de robustesse des barrières — c'est l'apport le plus mature, et c'est précisément ce que F1 décrit. *Deuxième étage* : l'aide à la lecture de ces indicateurs par dialogue d'interrogation — c'est ce que la fonction 2 décrit, et c'est l'usage le plus original que ces plateformes commencent à proposer. *Troisième étage* : la recommandation d'action en temps réel, voire la prescription automatisée, présentée comme le prolongement naturel des deux premiers. Cette gradation ressemble, vue du tableau de bord, à un raffinement fonctionnel. Vue de la doctrine de sécurité, elle franchit une frontière : les deux premiers étages outillent le jugement managérial, le troisième le remplace. Le pilotage de la robustesse en environnement sociotechnique complexe ne tolère pas cette substitution. C'est cette articulation — accepter les deux premiers étages, résister au troisième — qui reste à construire localement, organisation par organisation.

Tensions FOH principales

Sur la **compétence**. Un manager qui pilote son périmètre depuis un tableau de bord sophistiqué peut perdre, à terme, la compréhension du travail réel qui s'effectue sous sa responsabilité. Cette érosion est particulièrement insidieuse parce qu'elle se mesure mal — un manager peut être un excellent lecteur de tableau de bord et un piètre connaisseur de son terrain, sans que la dissociation soit perceptible par les indicateurs eux-mêmes. La parade tient à la préservation de moments de présence terrain non instrumentés, qui ne sont pas une régression, mais une condition de la qualité de la lecture des indicateurs.

Sur **l'automation complacency**. Le dialogue d'interrogation introduit par F2 produit un risque paradoxal : plus la qualité apparente du dialogue est élevée (réponses fluides, justifications cohérentes, navigation aisée dans la chaîne d'agrégation), plus le manager est conduit à attribuer à l'IA une fiabilité qu'elle n'a pas nécessairement. **[Parasuraman et Manzey 2010]** ont documenté ce mécanisme : passé un certain seuil de fiabilité apparente, l'opérateur cesse de vérifier, ne détecte plus les défaillances du système, et ne corrige pas ce déficit par l'entraînement. La parade ne tient pas à la formation à l'esprit critique — elle a été testée et ne suffit pas. Elle tient à l'exposition périodique à des situations où l'IA est manifestement défaillante, pour que le manager conserve une compétence opérationnelle de contestation, et à la pratique régulière d'analyses transversales conduites sans l'IA.

Sur le **blanchiment de légitimité**. Un Codir peut ratifier l'état de la robustesse tel que l'IA le rend lisible, sans interroger les paramètres qui produisent cette lecture — quelle variable est consolidée, quelle pondération est appliquée, quel seuil déclenche l'alerte. Le tableau de bord tient lieu de gouvernance. La parade exige une compétence exécutive nouvelle — lire un indicateur composite sans s'y soumettre, remonter à la chaîne d'agrégation, contester un seuil — qui ne s'improvise pas et qui demande une formation spécifique du management exécutif. La signature d'une décision sur la base d'un indicateur composite ne peut pas être ratifiée sur la base de l'indicateur : elle doit être délibérée à partir de l'indicateur.

Sur la **culture juste**. Les indicateurs descriptifs de robustesse — disponibilité d'une barrière, qualité du STOP, densité des précurseurs, écart entre prescrit et opérant — sont par nature plus granulaires qu'une recommandation agrégée, et donc plus facilement détournables en outil de surveillance individuelle. Un indicateur consolidé à l'échelle d'une équipe ou d'un poste peut être ressenti comme un outil de contrôle quelle que soit son intention initiale. Le silence remplace alors la déclaration sincère, et la matière qui alimentait le pilotage se dégrade. La parade tient à la séparation explicite des usages : ce qui sert à piloter la robustesse collective ne doit pas être réutilisé pour évaluer les individus, et cette séparation doit être contractualisée, vérifiable, et opposable.

Où placer le curseur

- **Déléguer** — la consolidation transverse des données (GMAO, SCADA, SIRH, SI HSE) et la production d'indicateurs descriptifs de robustesse des barrières — disponibilité conjointe des barrières critiques, densité des précurseurs, cohérence prescrit/opérant. C'est ici que l'IA offre la valeur la plus tangible et la moins discutable, à condition que l'indicateur reste descriptif — il rend visible une situation, il ne prescrit pas l'action à mener.
- **Assister** — le dialogue d'interrogation par lequel le manager rend lisible l'état de la robustesse : remonter la chaîne d'agrégation, contester une corrélation douteuse, croiser avec la connaissance du terrain. Niveau d'assistance *conseiller* exclusivement. Le niveau d'assistance *producteur* — recommandation d'action ou déclenchement automatique — est écarté pour le pilotage de la robustesse. Cette position est tenue par l'auteur malgré la pression du marché, parce que la décision en environnement sociotechnique complexe mobilise simultanément la technique, le système de management, les FOH et la dimension culturelle. L'IA peut outiller le jugement sur la première dimension ; elle ne peut pas le remplacer sur les autres.
- **Sanctuariser** — l'arbitrage stratégique en Codir. La décision qui découle de la lecture de la robustesse — modifier une barrière, réviser un équilibre du modèle, qualifier un mécanisme révélé par l'IA comme acceptable ou non — est un jugement de valeur exclusif au Codir : à délibérer à partir de l'indicateur, pas à ratifier sur la base de l'indicateur. Et l'audit périodique du modèle lui-même doit être inscrit comme une fonction de sécurité à

part entière, pas comme une tâche technique : c'est la condition pour que la cartographie qu'il produit reste cohérente avec le système qu'elle prétend décrire.

Et pour l'entreprise étendue ? Le pilotage par indicateurs des sous-traitants présente une difficulté spécifique. Un taux de fréquence affiché au vert peut coexister avec une sous-déclaration structurelle liée à la peur des pénalités contractuelles. Une IA qui consolide des indicateurs sans croiser ces données avec le REX réel des chantiers et l'état effectif des barrières automatise l'aveuglement. Et l'application d'indicateurs prescriptifs aux entreprises extérieures sans que celles-ci aient accès à la matière qui les produit — données, règles, modèle de sécurité du donneur d'ordre — institue une asymétrie de pilotage qui contrevient au principe d'exigence homogène posé tout au long de ce chapitre. Le pilotage de la robustesse partagée suppose une transparence partagée — c'est l'une des conditions sans laquelle les indicateurs reflètent la façade du contrat plutôt que la réalité du terrain.

Synthèse — Ce que les 6 questions ont en commun

Les six questions de ce chapitre partagent une même architecture. Dans chaque cas, l'IA traite un problème informationnel réel — données dispersées, non corrélées, enfouies — que les outils traditionnels ne savent pas résoudre. Et dans chaque cas, la qualité du résultat ne se joue pas au niveau de la performance technique, mais à celui de trois conditions organisationnelles que les déploiements ont tendance à sous-estimer :

- L'explicitation du modèle de sécurité : Sans modèle explicite, l'IA encode des arbitrages implicites, apprend des biais non documentés, et produit des résultats dont la cohérence avec la doctrine de l'organisation devient difficile à évaluer. Plusieurs apports introduits dans ce chapitre n'ont de sens que si cette condition est tenue : la matrice à deux axes — potentiel de gravité, potentiel d'apprentissage — suppose un modèle qui définit ce qu'apprendre veut dire pour l'organisation ; un premier jet d'analyse produit par l'IA suppose un référentiel d'analyse explicité ; un indicateur descriptif de robustesse encode lui-même une thèse sur ce qui compte — quelles variables sont consolidées, quels seuils signalent un affaiblissement. On peut formuler l'observation autrement : l'IA peut aider à mettre une doctrine à l'épreuve, elle ne l'invente pas.

  Cette première condition appelle un usage qui mérite d'être souligné parce qu'il revient à plusieurs reprises dans ce chapitre — en Q2 explicitement, en Q5 comme contribution propre, en Q6 par implication : **l'IA peut servir de miroir au modèle de sécurité** de l'organisation. En comparant ce qui est prévu, ce qui est compris et ce qui est fait, elle rend visibles des écarts que les outils traditionnels ne savaient pas toujours mettre en évidence. La fonction miroir n'est pas une fonctionnalité supplémentaire à programmer ; c'est un effet secondaire des dispositifs d'IA correctement gouvernés. Elle suppose qu'on accepte de regarder ce qu'elle révèle — et que ce qu'elle révèle puisse alimenter la révision du modèle plutôt que sa défense.

- Le **choix délibéré du niveau d'assistance** : La gradation du degré d'intervention de l'IA est universelle, mais elle prend des formes différentes selon la nature de la décision en jeu. Pour la préparation, la détection, la conduite en temps réel et l'apprentissage, elle se déploie sur le continuum *outil/conseiller/producteur* défini au chapitre 3. Pour le pilotage de la robustesse, où l'enjeu n'est pas tant qui produit la mesure que qui produit la décision qui en découle, elle prend la forme d'une distinction plus structurelle : *outillage de la lecture/substitution au jugement managérial*. L'IA peut légitimement rendre lisible un état complexe — c'est la fonction décisive en environnement sociotechnique complexe, où la lecture du paysage des indicateurs est elle-même un acte cognitif difficile. Elle ne peut pas pour autant produire la décision qui en découle : cette décision mobilise simultanément la technique, le système de management, les FOH et la dimension culturelle — quatre dimensions dont seule la première est traitable par les données structurées que l'IA ingère.

  Dans tous les cas, le principe est le même : plus l'IA fait, plus le risque d'atrophie des compétences et de blanchiment de légitimité s'accroît. Ce choix gagne à être explicite, formalisé et révisé périodiquement — plutôt que de s'installer par glissement progressif non gouverné. Et certains arbitrages — la qualification du double potentiel d'un événement, la décision de retenir un événement pour analyse approfondie, la décision tirée de la lecture de la robustesse — sont des actes de gouvernance partagée qui appellent une délibération en Codir : l'IA détecte, cote, rend lisible et pousse l'information ; l'organisation discute et arbitre.

- La **préservation délibérée des compétences humaines** : L'IA qui fonctionne bien en routine tend à rendre l'humain moins compétent pour les situations qui sortent du cadre — précisément celles où il est irremplaçable. On peut anticiper que cette érosion soit d'autant plus délicate à compenser qu'elle ne se manifeste qu'au moment où le système défaille, c'est-à-dire au moment où la compétence aurait justement été utile. Et le risque ne se limite pas à l'atrophie lente : un dialogue d'explicabilité de bonne qualité peut produire une *illusion de compréhension* qui s'installe immédiatement, indépendamment de la fiabilité réelle du système.

Trois conditions reviennent dans les six questions :

- Maintenir des espaces d'exercice du jugement autonome, par des activités conduites sans assistance à intervalles réguliers ;
- Concevoir l'IA en mode socratique — qui interroge plutôt qu'elle ne prescrit — quand la fonction le permet ; c'est l'apport central de Fonction 2 en Q5, où l'IA n'est pas là pour *trouver* les causes, mais pour *empêcher l'analyste de s'arrêter trop tôt* ;
- Protéger institutionnellement la capacité de contestation du score algorithmique ou de la lecture algorithmique, en faisant en sorte que s'écarter de la sortie de l'IA ne soit pas plus coûteux que la suivre.

Ces trois conditions ne sont pas des précautions de déploiement : elles relèvent de la conception. Traiter les FOH comme une couche d'accompagnement du changement, ajoutée après que la solution technique a été construite, expose à un échec documenté : les REX sur les modifications d'installations montrent que cette approche séquentielle — d'abord la technique, ensuite l'humain — produit des systèmes techniquement performants, mais opérationnellement rejetés ou détournés. Pour l'IA en sécurité, l'enjeu est plus aigu encore : un système d'aide à la décision dont les modes dégradés n'ont pas été pensés à la conception ne sera pas corrigeable par une formation. Les questions FOH de ce chapitre méritent donc d'être lues comme une grille de spécifications à intégrer en amont, pas comme un commentaire a posteriori sur des solutions déjà arrêtées.

Un fil transversal traverse les six questions : **l'entreprise étendue est systématiquement le maillon le plus fragile**. Les données des sous-traitants sont les plus cloisonnées, le silence organisationnel y est le plus puissant, et la gouvernance partagée des données la moins construite. À chaque question correspond une exigence de symétrie qu'il faut tenir — équité d'accès au niveau d'assistance, accessibilité du canal humain de remontée, partage du résultat des analyses transversales. Le chapitre 6 développera cette dimension.

Point clé

L'IA ne remplace ni le modèle de sécurité, ni le jugement humain, ni la relation de confiance qui rend l'apprentissage possible. Elle peut les renforcer ou les fragiliser, et l'expérience invite à anticiper que ce qui fait la différence n'est pas la technologie déployée — c'est la manière dont l'organisation décide de l'utiliser, en connaissance de cause et en cohérence avec son modèle de sécurité.

5

L'angle mort de l'entreprise étendue

Tout au long du chapitre 4, un encadré récurrent a signalé, question après question, que l'entreprise étendue est le maillon le plus fragile du déploiement de l'IA en sécurité industrielle. Les données des sous-traitants sont les plus cloisonnées, le silence organisationnel y est le plus puissant, la gouvernance partagée des données est la moins construite. Ce chapitre développe ce que les encadrés ne pouvaient que signaler : pourquoi le problème est relationnel avant d'être technologique, quel préalable organisationnel est nécessaire avant tout déploiement d'IA, et à quelles conditions l'IA peut renforcer la sécurité de l'entreprise étendue au lieu d'amplifier ses dysfonctionnements.

Le constat de départ est simple : la sous-traitance représente une grande partie des heures travaillées sur des sites industriels à risque, et une proportion comparable des accidents graves. Pourtant, la quasi-totalité des publications sur l'IA en industrie raisonnent implicitement dans le périmètre du donneur d'ordre, comme si les données étaient entièrement sous son contrôle. C'est un angle mort considérable.

5.1 Le problème est relationnel, pas technologique

Dans la configuration typique, les échanges de données entre donneur d'ordre et entreprise extérieure suivent un schéma asymétrique. Le donneur d'ordre transmet les cahiers des charges et pièces techniques, les plans de prévention, les consignes spécifiques au site — généralement en PDF. L'entreprise extérieure transmet ses habilitations, sa qualification MASE ou équivalent, ses statistiques de sécurité, ses procédures et modes opératoires, ses comptes rendus d'intervention — également en PDF ou formulaires papier. Chaque partie reste propriétaire de ses données dans son propre système d'information. Le croisement est manuel, partiel et non reproductible.

Cette situation n'est pas un retard technique à rattraper. C'est le reflet fidèle **d'une relation structurellement asymétrique**, illustrée à la figure 5.1. Le donneur d'ordre prescrit, contrôle et sanctionne. L'entreprise extérieure exécute, rend compte et dépend économiquement du renouvellement du contrat. Dans cette configuration, la transparence complète n'est pas un idéal à atteindre — c'est un risque pour le sous-traitant. Comme le montre la fragilité 4 du chapitre 1, le silence organisationnel aux interfaces de l'entreprise étendue n'est pas un choix individuel : c'est un comportement rationnel dans un système où signaler un problème peut coûter un contrat.

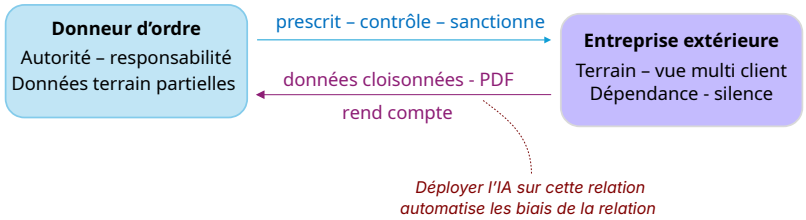

Fig. 5.1 La relation asymétrique entre donneur d'ordre et sous-traitant. Les données que le sous-traitant communique au donneur d'ordre sont soumises à différents biais, auxquels il faut être attentif avant de réaliser un traitement automatisé.

Déployer de l'IA sur ces fondations ne produit pas de l'intelligence — cela automatise les biais de la relation. Un algorithme de cotation du niveau de risque à maîtriser qui utilise les statistiques de sécurité des prestataires pour attribuer les marchés crée un cercle vicieux : les prestataires ayant moins d'événements déclarés sont favorisés, ce qui incite à la sous-déclaration. L'IA amplifie alors un biais culturel existant au lieu de le corriger.

Un modèle de maintenance prédictive entraîné uniquement sur les données du donneur d'ordre ne voit qu'une partie de l'histoire d'un équipement si la maintenance est sous-traitée. Inversement, le prestataire dispose d'une vision multi-clients que le donneur d'ordre n'a pas — mais qu'il ne partage pas, parce que c'est sa valeur ajoutée commerciale.

Point clé

Avant de déployer de l'IA sur les activités sous-traitées, la question préalable n'est pas technique, mais relationnelle. Quelle souveraineté partagée sur les données est acceptable pour les deux parties ? Quelle gouvernance garantit que l'IA ne devient pas un outil de contrôle asymétrique du donneur d'ordre sur l'entreprise extérieure ? C'est une question de culture de sécurité de l'entreprise étendue, pas de technologie.

5.2 L'observatoire comme préalable — pas comme option

Le cahier *Donneur d'ordre, architecte de la culture de sécurité* **[Quiot 2026]** propose un dispositif qui, bien que conçu indépendamment de la question IA, en constitue à la fois le préalable et l'infrastructure permanente : **l'observatoire sectoriel de la culture de sécurité**. L'idée est de construire, à l'échelle d'un segment industriel ou d'un cluster d'entreprises, un cadre partagé d'apprentissage qui fournit une ligne de base commune, suit des tendances quantitatives et qualitatives, et alimente des décisions partagées.

Pour le déploiement IA, l'observatoire ne se réduit pas à une étape qu'on coche avant de passer à autre chose — il devient l'institution permanente qui doit grandir à mesure que l'IA grandit. Plus l'IA agit en substitut de l'effort humain là où la matière apprenante est la plus informelle et qualitative — REX, observations terrain, analyses causales — et dont la qualité dépend en partie du temps qu'on y consacre, plus l'observatoire devient critique pour maintenir la production de connaissance générale. C'est un retournement de perspective qu'il faut nommer explicitement : l'observatoire n'est pas un coût ponctuel à amortir avant de récolter les bénéfices du déploiement de l'IA, c'est un investissement structurel dont l'ampleur doit suivre l'ambition IA.

La pertinence de cet observatoire pour l'IA tient à trois raisons.

- **L'observatoire construit la confiance nécessaire au partage de données**. Un sous-traitant ne partagera ses données de sécurité avec son donneur d'ordre que s'il a la certitude que ces données ne seront pas utilisées contre lui — pour le sanctionner, le déréférencer ou réduire ses marges commerciales. L'observatoire, par sa gouvernance paritaire, sa charte d'usage centrée sur l'amélioration et ses règles d'anonymisation, crée un espace où la transparence est sûre. Sans cet espace, aucun algorithme ne disposera de données fiables sur l'entreprise étendue.
- **L'observatoire produit les données structurées dont l'IA a besoin**. Tant que les échanges entre donneur d'ordre et entreprises extérieures se font en PDF, le traitement analytique est impossible. L'observatoire, en imposant un socle commun d'indicateurs comparables et des formats d'échange structurés, crée les conditions techniques du croisement de données. Ce n'est pas un projet informatique : c'est un travail de normalisation des référentiels entre partenaires, qui suppose un accord sur ce qu'on mesure, comment on le mesure et pourquoi.
- **L'observatoire est le lieu où les désaccords sur l'usage de l'IA peuvent être traités**. Qui décide des algorithmes de cotation du niveau de risque à maîtriser ? Le sous-traitant voit-il les mêmes résultats que le donneur d'ordre ? Que se passe-t-il si la cotation est contestée ? Ces questions ne peuvent pas être réglées dans un contrat bilatéral où le rapport de force est déséquilibré. Elles nécessitent un espace tiers, avec un tiers de confiance garant de l'impartialité. L'observatoire est cet espace.

5.3 Les conditions d'un déploiement qui ne renforce pas l'asymétrie

Si l'observatoire construit le cadre de confiance, le déploiement de l'IA dans le contexte de l'entreprise étendue doit respecter des conditions spécifiques qui dépassent celles du chapitre 4.

5.3.1 La réciprocité dans l'accès aux données

Si le donneur d'ordre accède aux données du prestataire pour alimenter une cotation du niveau de risque à maîtriser ou une analyse transversale, le prestataire doit avoir accès aux données qui le concernent dans le système du donneur d'ordre. Le REX du donneur d'ordre sur les interventions similaires, l'historique des équipements sur lesquels le prestataire va intervenir, les alertes issues de la supervision instrumentée : ces informations sont précieuses pour la sécurité du prestataire et de ses équipes. Les retenir, c'est priver le sous-traitant des moyens de se protéger. Le partage doit être bidirectionnel, ou il ne sera pas accepté.

5.3.2 Le cas particulier de la supervision temps réel

La réciprocité dans l'accès aux données prend une forme opérationnelle spécifique quand le dispositif est porté par les intervenants eux-mêmes — capteurs communicants, balises environnementales, caméras de vérification de gestes critiques. Dans ces situations, la règle générale de réciprocité se traduit par un principe opérationnel : les salariés d'une entreprise extérieure sont supervisés en première intention par un salarié de leur propre entreprise, pas par le donneur d'ordre. Le DO reçoit l'information qui le concerne au titre de son rôle d'exploitant — consolidée, différée, anonymisée selon ce que la situation exige — mais il n'exerce pas la supervision individuelle directe à la place de l'employeur réel.

Ce principe, éprouvé dans les générations pré-IA de dispositifs de supervision, n'est pas un raffinement procédural : il préserve l'espace relationnel interne à l'entreprise extérieure et évite que la supervision technique ne devienne un outil d'asymétrie de contrôle.

Les configurations où l'entreprise extérieure ne peut pas déployer un superviseur propre — petites structures, interventions ponctuelles, rangs 2 et 3 de sous-traitance — appellent des montages spécifiques qui préservent le principe sans en figer la forme : superviseur mutualisé au niveau d'une branche professionnelle, tiers de confiance indépendant mandaté par l'ensemble des sous-traitants d'un site, cellule de supervision inter-entreprises sur les opérations de grande ampleur. L'essentiel est que la lecture directe des données individuelles ne relève jamais, par défaut, du donneur d'ordre.

La déclinaison opérationnelle de ce principe — finalité déclarée, scope fonctionnel borné, gouvernance de la lecture, mécanismes antidérive dans la durée — est traitée en Q4 du chapitre précédent.

5.3.3 La gouvernance des algorithmes

Qui définit les critères de cotation du niveau de risque à maîtriser ? Qui valide les données d'entraînement ? Qui décide des seuils ? Si ces choix sont faits unilatéralement par le donneur d'ordre, le sous-traitant subit un système dont il ne maîtrise ni les règles ni les conséquences.

La gouvernance des algorithmes utilisés dans la relation donneur d'ordre / sous-traitant doit être partagée — ou au minimum transparente. Le sous-traitant doit savoir sur quels critères il est évalué, pouvoir contester un résultat qu'il estime erroné, et disposer d'un recours auprès du tiers de confiance de l'observatoire.

5.3.4 La protection contre les usages détournés

Les données partagées dans le cadre de l'observatoire ou du déploiement d'IA doivent être protégées contre trois détournements.

- Le premier est l'utilisation dans les appels d'offres : si les données de sécurité d'un prestataire sont utilisées pour le déréférencer sans que les conditions imposées par le donneur d'ordre soient prises en compte, le système est injuste et le prestataire cessera de partager.
- Le deuxième est l'utilisation dans les contentieux : si les données partagées pour l'apprentissage sont utilisées dans un litige, la confiance est détruite.
- Le troisième est la diffusion non contrôlée : les données d'un prestataire chez un client ne doivent pas être accessibles à ses concurrents sans anonymisation.

5.3.5 Le diagnostic de culture de sécurité comme préalable, pas la cotation

Un diagnostic de culture de sécurité de l'entreprise étendue est un préalable nécessaire avant tout déploiement d'IA dans le contexte DO/ST. Mais ce diagnostic doit être conçu pour neutraliser les biais structurels de la relation : le biais de survie économique qui pousse à donner les « bonnes réponses » plutôt que les vraies, le biais de désirabilité sociale, le biais d'attribution asymétrique où chaque niveau rejette la responsabilité sur l'autre.

Point d'attention : Un diagnostic de culture de sécurité de l'entreprise étendue qui ne neutralise pas les biais de la relation — survie économique, désirabilité sociale, attribution asymétrique — ne mesure pas la réalité : il mesure ce que chaque partie a intérêt à montrer. Il produit une image « pastèque », verte en surface et rouge à l'intérieur, qui donne l'illusion de la maîtrise et endort la vigilance. Une IA branchée sur cette matière n'amplifie pas l'apprentissage : elle automatise l'aveuglement.

5.4 Trois pistes pour avancer

Trois approches émergent dans la littérature et dans la pratique des industriels les plus avancés, sans qu'aucune soit encore dominante.

- **Les référentiels de données partagés**. Les structures de prévention de branche — qu'il s'agisse d'organismes paritaires (OPPBTP…), de syndicats professionnels (France Chimie, UFIP), d'opérateurs publics tiers de confiance (BARPI…) — pourraient évoluer vers des plateformes de partage de données de sécurité anonymisées et agrégées, exploitables par l'IA sans exposer les données individuelles de chaque prestataire. Aucune de ces structures n'est aujourd'hui taillée pour ce rôle, et la question structurante est moins celle du porteur que celle de la gouvernance : qui décide des règles d'accès, qui garantit l'anonymisation, qui arbitre les litiges, qui en assure le financement permanent. Le projet Data4NuclearX (consortium EDF/CEA/Sopra Steria) explore cette voie dans le nucléaire avec un déploiement prévu en 2028. La clé est la gouvernance : qui décide des règles d'accès, qui garantit l'anonymisation, qui arbitre les litiges.
- **Les jumeaux numériques de chantier partagés.** Un modèle de risque commun, alimenté par le donneur d'ordre et les entreprises extérieures pour la durée d'une intervention, pourrait offrir à tous les acteurs la même vision de l'état du chantier, des coactivités et des risques. C'est techniquement faisable ; la difficulté est organisationnelle — cela suppose que chaque partie accepte de rendre ses données visibles en temps réel — ce qui reste un obstacle majeur dans la plupart des configurations actuelles.
- **L'apprentissage fédéré**. Cette technique, formalisée par **[McMahan et al. 2017]**, permet d'entraîner un modèle de **machine learning** sur les données de plusieurs entreprises sans que les données brutes ne quittent chaque système d'information. Chaque participant entraîne le modèle localement et partage les paramètres appris, pas les données. L'approche est prometteuse pour construire des modèles prédictifs à l'échelle d'un secteur — par exemple un modèle de maintenance prédictive entraîné sur l'expérience collective de plusieurs exploitants et prestataires. Elle commence à faire l'objet de déploiements industriels en maintenance prédictive et inspection qualité **[Pruckovskaja et al. 2023]**, mais reste au stade de recherche pour les usages directement liés à la prévention des événements graves, où l'hétérogénéité des données entre sites et le caractère rare des signaux à apprendre posent des difficultés méthodologiques non résolues.

5.5 Ce que ce chapitre dit au dirigeant

L'entreprise étendue est l'endroit où la thèse du chapitre 3 est la plus tendue. Expliciter le modèle de sécurité suppose que les deux parties — donneur d'ordre et entreprise extérieure — partagent la même compréhension de ce qui protège, de ce qui est négociable et de ce qui ne l'est pas.

Choisir ce que l'on délègue, assiste ou sanctuarise suppose un accord sur qui décide, avec quelles données et sous quelle responsabilité. Utiliser l'IA comme miroir suppose que les deux parties acceptent de regarder le reflet — y compris quand il montre des dysfonctionnements dont l'une ou l'autre est responsable. Rien de tout cela n'est possible sans un travail préalable de construction de la confiance. Ce travail est lent, incrémental et fragile. Il ne peut pas être accéléré par la technologie — il doit précéder la technologie. L'observatoire évoque au § 5.2

n'est pas un dispositif complémentaire au déploiement de l'IA dans l'entreprise étendue : c'est sa condition de possibilité.

À l'échelle de la filière nucléaire française, le programme Data4NuclearX est une traduction industrielle de cette intuition : avant les algorithmes, les langages communs ; avant la cotation du niveau de risque à maîtriser, la souveraineté partagée sur les données. L'articulation entre l'observatoire (niveau sectoriel) et Data4NuclearX (niveau de filière) dessine l'infrastructure de confiance qui précède toute IA en sécurité dans l'entreprise étendue.

Question ouverte

Les industries à haut risque peuvent-elles construire un espace de partage de données de sécurité entre donneurs d'ordre et entreprises extérieures qui soit à la fois assez transparent pour alimenter l'IA, assez protégé pour préserver la confiance et assez équitable pour ne pas reproduire sous forme algorithmique l'asymétrie de la relation contractuelle ?

5.5.1 Arbitrer le couple performance/souveraineté avant le déploiement

Trois questions doivent être tranchées en Codir, et non dans l'outillage technique *a posteriori* :

- ▷ Sur quelles familles de données accepte-t-on de recourir à des LLM hébergés hors juridiction nationale ?
- ▷ Quelle doctrine d'usage des outils grand public par les collaborateurs ?
- ▷ Quelle part des usages relève d'un déploiement souverain, même au prix d'un écart de performance ?

On peut anticiper que les organisations qui ne tranchent pas ces questions en amont les verront ressurgir en crise — à la première fuite documentée, à la première demande d'accès d'une autorité étrangère, ou à la première polémique publique. La confidentialité des données de sécurité n'est pas un sujet technique ; c'est un sujet de gouvernance.

5.5.2 Exiger que le FOH soit traité dans la spécification, pas dans le correctif

Un projet IA en sécurité dont le cahier des charges ne comporte pas d'analyse explicite des impacts sur les compétences, sur la vigilance, sur le sensemaking en situation dégradée, sur la chaîne de responsabilité et sur la culture de sécurité n'est pas un projet à améliorer — c'est un projet à réécrire. L'expérience des modifications d'installations industrielles le documente de manière constante : ce qui n'est pas spécifié en amont coûte dix fois plus cher en aval, quand il reste récupérable.

Le comité de direction joue ici un rôle que personne d'autre ne peut tenir : exiger que les livrables de spécification comportent une analyse FOH au même titre qu'une analyse de sûreté fonctionnelle ou de cybersécurité, et refuser les dossiers qui ne la comportent pas. C'est une exigence de méthode, pas de culture.

6

Quelques bonnes questions à se poser pour un Codir

Ce chapitre ne résume pas les précédents. Il les traduit en sept questions qu'un Codir devrait avoir traitées avant de considérer que son organisation est prête à déployer l'IA. Chaque question est accompagnée d'un test de maturité.

Le déploiement de systèmes IA dans une organisation à haut risque, n'est pas l'installation d'un logiciel supplémentaire. **On ne déploie pas une IA — on change la manière de travailler**. Conformément à la mosaïque hétérogène posée au chapitre 2, ce que recouvre concrètement un tel déploiement n'est jamais une seule application : c'est une combinaison progressive de modules, parfois embarqués dans des outils existants, parfois apportés sous forme d'applications dédiées, qui s'insèrent dans des collectifs et reconfigurent leurs façons de faire.

Cette nature progressive a une conséquence directe : il n'y a pas de grand soir IA. La logique qui s'impose est celle d'un **déploiement par paliers** — poser d'abord un cadre et une architecture technique, puis ouvrir une fonction dans un collectif, en maîtriser les impacts FOH avant de la dupliquer dans un collectif suivant, comme dans la conduite du changement "lean". À chaque palier, l'organisation apprend autant qu'elle déploie. C'est ce caractère cumulatif qui justifie la photographie de départ : on ne transforme pas ce qu'on ne connaît pas, et on ne maîtrise pas la trajectoire d'un déploiement par paliers sans connaître le point de départ.

6.1 Q1 : Avant de se lancer, avons-nous fait la photographie de départ ?

Tout programme de transformation suppose un diagnostic préalable de l'état réel. Ce préalable est antérieur aux six questions qui suivent : il en conditionne la pertinence même. Pour l'IA, ce diagnostic doit répondre à une question simple, mais difficile. Dans le cadre de la sécurité industrielle, l'organisation présente-t-elle aujourd'hui une fissure stable que l'IA peut aider à colmater, une fissure qui s'élargit silencieusement que l'IA pourrait accélérer, ou une brèche que l'IA aggraverait quoi qu'on fasse ? Cette distinction n'est pas culturelle, elle est paramétrique : elle détermine si le déploiement va combler ou élargir.

Point d'attention : Une organisation qui a des brèches est précisément celle qui ne sait pas qu'elle en a : une brèche ne se voit pas de l'intérieur — par déni, par érosion de la capacité à voir, ou par silence organisationnel. Le diagnostic préalable au déploiement ne peut donc pas être autoadministré. Demander à l'organisation d'évaluer elle-même si elle est prête, c'est lui demander de voir ce que sa situation l'empêche justement de voir.

Dans le cadre des transformations de cultures de sécurité, la pratique est de réaliser un diagnostic, fondé sur des entretiens, des observations terrain et une analyse documentaire, qui produit la photographie de départ avant tout programme de transformation. Pour le déploiement IA, elle doit être complétée par trois modules spécifiques :

1. **La maturité informationnelle de l'organisation**. Quel est l'état réel des données qui alimenteront l'IA ? Les bases REX sont-elles structurées et vivantes, ou archivées et figées ? Les comptes rendus narratifs sont-ils rédigés honnêtement ou pour rassurer ? Les sources hétérogènes restent-elles isolées dans leurs silos ? La qualité des remontées humaines (climat de signalement) ne suffit pas à garantir la qualité des données structurées qui en résultent.

2. **La maturité du modèle de sécurité**. L'organisation a-t-elle explicité son modèle de sécurité au sens du chapitre 3 — arbitrages structurants, barrières critiques nommées, principes directeurs formulés ? Ou fonctionne-t-elle avec un modèle implicite que personne n'a écrit ? C'est le modèle formalisé qui sera enseigné à la machine, et son absence est une vulnérabilité spécifique au déploiement IA.
3. **La cartographie des zones fragiles**. Sur les domaines PAGEM où la production de connaissance dépend le plus de l'engagement humain — REX, analyses causales, observations terrain —, où sont aujourd'hui les compétences fortes et où sont les compétences fragiles ? Quelles zones sont protégées par la culture, quelles zones sont déjà en érosion silencieuse ? C'est cette distribution qui détermine où le déploiement va combler une fissure et où il va élargir une brèche.

L'ampleur du diagnostic doit être proportionnée à la criticité du déploiement envisagé. Pour un cas d'usage en niveau d'assistance *outil* sur une fonction non critique, un diagnostic léger peut suffire. Pour un déploiement en niveau d'assistance *conseiller* ou *producteur* sur une fonction de sécurité, le diagnostic complet devient non négociable. Et il ne s'agit pas d'un acte ponctuel : la maturité bouge, la phase de transition modifie elle-même les paramètres qu'elle traverse, et le diagnostic préalable doit s'inscrire dans une logique d'observation continue, articulée avec l'observatoire de la culture de sécurité.

Note d'orientation : La photographie de départ ne porte pas seulement sur les fragilités PAGEM ; elle porte aussi sur la gouvernance pressentie du déploiement. Une indication tirée de l'expérience : un déploiement IA qui touche à la sécurité ne peut pas être porté par les seules fonctions HSE, sûreté ou DSI. Les impacts traversent tous les domaines de performance de l'organisation et engagent en profondeur les compétences, les marges de manœuvre et la présence terrain — autant de dimensions FOH qui appellent un portage managérial de direction. Le détail de l'arbitrage entre coût et bénéfice du déploiement, et les modalités précises du financement, dépassent l'objet de ce document : ils dépendent du contexte et seront évoqués comme tension assumée au chapitre 7.

Test de maturité

- ▷ Avez-vous, dans les années qui précèdent votre projet IA, conduit un diagnostic de votre culture de sécurité ?
- ▷ Sur les domaines de la PAGEM, savez-vous distinguer ceux qui sont aujourd'hui des fissures stables, ceux qui s'élargissent silencieusement, et ceux qui sont déjà des brèches ?
- ▷ Avez-vous identifié les zones de votre organisation où la qualité d'effort cognitif humain est la plus fragile aujourd'hui — celles où l'IA aurait l'effet de substitution le plus rapide ?

6.2 Q2 : Notre modèle de sécurité est-il explicite ?

Avant de confier quoi que ce soit à une IA, il faut savoir ce qu'on protège et comment. Un modèle de sécurité explicite n'est pas un document de plus dans le système de management : c'est la réponse partagée à la question « *qu'est-ce qui fait qu'il n'y aura pas d'accident grave dans notre organisation ?* »

Si cette réponse n'est pas formulée explicitement, l'organisation ne sait pas ce qu'elle enseignera à la machine. Les données d'entraînement encoderont des arbitrages implicites, des biais non documentés et des croyances que personne n'a validées. Ce travail d'explicitation a une vertu propre, indépendante de l'IA : réduire les écarts entre ce qui est prévu, ce qui est compris et porté par le management et ce qui est fait rend l'organisation plus sûre — avec ou sans IA.

Test de maturité

- ▷ Pouvez-vous nommer vos cinq barrières critiques sans consulter un document ?
- ▷ Vos managers de terrain donnent-ils la même réponse que votre directeur HSE à la question « qu'est-ce qui nous protège d'un accident grave ? » ?
- ▷ Si vous demandiez à trois préparateurs quels arbitrages entre production et sécurité ils font quotidiennement, leurs réponses vous surprendraient-elles ?

6.3 Q3 : Nos données disent-elles la vérité ?

Un algorithme ne vaut que ce que valent ses données. La question n'est pas *« avons-nous assez de données ? »*, mais *« nos données reflètent-elles la réalité du terrain ? »*. Si les quasi-accidents sont sous-déclarés, si les sous-traitants filtrent leurs remontées, si les rapports d'analyse sont rédigés pour rassurer plutôt que pour comprendre, alors l'IA entraînée sur ces données n'apprendra pas la réalité — elle apprendra les biais de votre culture de remontée. La qualité des données n'est pas un problème de système d'information. C'est un indicateur de maturité organisationnelle.

Test de maturité

- ▷ Votre taux de quasi-accidents déclarés a-t-il augmenté ces trois dernières années — ou a-t-il diminué parce que déclarer coûte plus que se taire ?
- ▷ Les entreprises extérieures déclarent-elles autant d'événements proportionnellement que vos équipes internes ?
- ▷ Vos bases de REX, votre GMAO, votre SIRH et votre supervision sont-ils dans des systèmes qui peuvent se parler — ou faut-il un stagiaire pour les croiser ?

6.4 Q4 : Qu'est-ce que nous confions à l'IA — et qu'en dit notre régulateur ?

La grille *déléguer, assister, sanctuariser* n'est pas un exercice théorique. C'est une décision de gouvernance. Pour chaque cas d'usage envisagé, le Codir doit choisir explicitement : est-ce que l'IA fait seule (déléguer), propose et l'humain décide (assister), ou est-ce que cette décision reste entièrement humaine (sanctuariser) ? Et à l'intérieur de « assister », le choix du niveau d'assistance — l'humain pilote l'IA, l'IA conseille l'humain, ou l'IA produit et l'humain supervise — change radicalement les impacts sur les compétences, la vigilance et la responsabilité. Ce positionnement ne devrait pas être un glissement progressif. Il devrait être formalisé et révisé.

Dès qu'un cas d'usage touche à un équipement classé, à une barrière de sûreté ou à une fonction réglementée, le Codir devrait engager le dialogue avec son autorité de contrôle — tôt, avant d'investir. Les régulateurs nucléaires internationaux (CNSC, ONR, NRC) ont publié en 2024 un document de principes qui pose que l'exigence de qualification doit être calibrée sur les conséquences d'une défaillance et le degré d'autonomie de l'IA (*cf.* l'annexe D). Mais aucun cadre stabilisé n'a encore émergé : le dialogue précoce permet de coconstruire les critères plutôt que de les subir.

Test de maturité

- ▷ Pour chaque projet IA en cours ou envisagé, pouvez-vous dire en une phrase s'il est en régime déléguer, assister ou sanctuariser ?
- ▷ Votre autorité de contrôle est-elle informée de vos projets IA touchant à la sécurité ?
- ▷ Si un score IA sous-évalue un risque et qu'un incident survient, qui dans votre organisation porte la responsabilité — et cette réponse est-elle formalisée ?

6.5 Q5 : Avons-nous pensé l'humain dès la conception — et construit l'espace de contestation ?

L'expérience des modifications d'installations industrielles montre que les FOH doivent être traités dans la spécification, pas dans le correctif. Pour l'IA, l'enjeu est plus aigu encore : un système d'aide à la décision dont les modes dégradés n'ont pas été pensés à la conception ne sera pas corrigeable par une formation. Le cahier des charges d'un projet IA en sécurité doit répondre à des questions avant la première ligne de code :

1. Quelle compétence humaine sera affectée ?
2. Quel impact sur la vigilance ?
3. Comment le sensemaking collectif sera-t-il préservé en situation dégradée ?
4. Qui sera responsable de quoi ?

5. Que se passe-t-il quand l'IA est indisponible ?

Mais l'analogie avec la modification d'installation a une limite : une vanne, une fois modifiée, a un comportement stable ; un système d'IA continue d'évoluer après sa mise en service. L'exigence FOH a donc deux temps — la conception, puis la durée — et porte sur deux dérives distinctes : celle du modèle, dont les performances se dégradent silencieusement, et celle de l'interaction humain-machine, où les compétences s'érodent et la signature devient rituelle. La première relève de la surveillance technique du modèle ; la seconde est l'objet des tests de viabilité du chapitre 3 et des indicateurs de trajectoire du chapitre 7.

Et la conception ne suffit pas sans un mécanisme vivant de contestation. Il faut désigner formellement les acteurs qui ont le droit et le devoir de dire « ce score ne reflète pas ce que je vois sur le terrain » — et les protéger institutionnellement dans ce rôle. Ce sont les praticiens de terrain, les préventeurs, les managers de proximité : leur légitimité vient de leur connaissance du travail réel, pas de leur compréhension de l'algorithme. La contestation d'un score IA doit fonctionner comme le droit de STOP : le doute profite au contestataire, pas à l'algorithme.

Point d'attention : L'espace de contestation est un dispositif prescrit — rôles désignés, protection institutionnelle, droit formalisé. Mais un droit formel ne devient un acte réel que si la culture juste et équitable est effectivement tenue : contester un résultat de l'IA et se tromper ne doit pas exposer à davantage qu'une erreur de bonne foi. Sans ce substrat, chacun comprend vite qu'il est moins coûteux de suivre l'algorithme que de s'en écarter, et le dispositif reste lettre morte. La culture juste en est la condition première — non la condition suffisante : encore faut-il que le coût de l'écart ne dépasse pas celui du suivi.

On conteste d'abord, on analyse ensuite. Sans cette capacité de contestation, le score algorithmique devient progressivement la référence partagée, le débat sur le risque réel se déplace vers un débat sur les métriques, et la signature humaine devient un blanchiment de légitimité.

Test de maturité

- ▷ Le cahier des charges de votre dernier projet IA comportait-il une analyse des impacts FOH ?
- ▷ Avez-vous désigné formellement qui a le droit de contester un score IA dans votre organisation ?
- ▷ Un opérateur qui conteste un score IA et qui a tort subit-il une conséquence ?
- ▷ Vos data scientists considèrent-ils les contestations terrain comme une source d'amélioration — ou comme un problème à gérer ?

6.6 Q6 : Par quoi commençons-nous — et à quoi renonçons-nous ?

L'IA ne s'ajoute pas aux priorités existantes. Elle les redistribue. Choisir un premier cas d'usage, c'est aussi choisir ce qu'on ne fera pas — et libérer les ressources pour le faire bien plutôt que de faire beaucoup mal. Deux pièges sont à éviter : le projet phare trop ambitieux qui échoue et discrédite l'IA pour cinq ans, et le saupoudrage de petits projets qui n'atteignent jamais la masse critique. Quatre critères pour sélectionner le premier cas d'usage :

- ▷ La **valeur ajoutée visible** — le résultat doit être perceptible par ceux qui travaillent, pas seulement par ceux qui pilotent.
- ▷ Les **données disponibles** — commencer par les gisements qui existent déjà, pas par ceux qu'il faut construire.
- ▷ Le **risque FOH maîtrisable** — privilégier les cas où le niveau d'assistance *outil* ou *conseiller* est naturel, pas ceux qui nécessitent d'emblée le niveau d'assistance *producteur*.
- ▷ Et la **réversibilité** — pouvoir revenir en arrière si le résultat déçoit ou si les effets FOH sont plus importants que prévu.

Il est certainement pertinent de commencer par les cas d'usage qui ne nécessitent pas de qualification réglementaire complexe. La recherche sémantique dans le REX, l'analyse de tendances dans les données de maintenance, la détection de redondances dans le référentiel documentaire — ces cas d'usage sont en niveau d'assistance *outil* ou *assister-conseiller*, ne

touchent pas directement aux barrières de sûreté, et produisent des résultats visibles rapidement. Ils construisent l'expérience, la compétence interne et la crédibilité nécessaires avant d'aborder les cas critiques.

Point d'attention : La détection de redondances ou la simplification documentaire ne sont « à faible risque » qu'à une condition : distinguer la complexité-opacité — documents contradictoires, jamais consultés, dont personne ne tient la vue d'ensemble — de la complexité qui travaille, dont la lourdeur impose l'effort de croisement et de reconstitution qui entretient la vigilance. Réduire la première est un gain ; fluidifier la seconde échange un problème d'opacité contre une érosion silencieuse de la vigilance. Le caractère « faible risque » d'un cas d'usage documentaire se juge à cette distinction, pas au volume traité.

Test de maturité

- ▷ Pouvez-vous nommer le premier cas d'usage IA que vous déploierez en sécurité — et expliquer pourquoi celui-là plutôt qu'un autre ?
- ▷ Quelles ressources libérez-vous pour ce projet — et à quoi renoncez-vous pour les libérer ?
- ▷ Si le projet échoue, pouvez-vous revenir en arrière sans dégradation de la sécurité ?

6.7 Q7 : Avons-nous pensé la gestion de la phase de transition ?

Les questions précédentes ont traité ce qu'un comité de direction doit avoir réglé *avant* de déployer. Celle-ci traite ce qu'il doit avoir réglé pour *traverser* le déploiement. Comme indiqué au chapitre 3, le déploiement de l'IA en sécurité industrielle n'est pas une décision ponctuelle — c'est une trajectoire qui s'étale sur plusieurs années et dont les premiers mois appellent des dispositifs particuliers. L'erreur stratégique n'est pas le mauvais choix de cas d'usage ; c'est de déployer comme on installe un logiciel classique, en supposant que le régime de croisière suivra mécaniquement la mise en service. Trois dimensions structurent cette phase de transition :

- ▷ **La survigilance transitoire**. Pendant la période qui suit un déploiement, les mécanismes de confiance humain-IA ne sont pas encore stabilisés. L'opérateur ne sait pas encore quand faire confiance à la recommandation de la machine et quand la contester. L'encadrement n'a pas encore calibré sa propre lecture des signaux que l'IA remonte.

  La littérature longitudinale sur la confiance humain-automatisation **[Hoff et Bashir 2015 ; de Visser et al. 2020]** montre que cette calibration se mesure en mois — voire en un à deux ans quand les situations significatives sont rares. La durée n'est pas paramétrable à l'avance : elle dépend de la fréquence d'usage, de la rareté des événements significatifs, de la diversité des situations rencontrées et de la qualité des feedbacks lorsque le système se trompe.

  Pendant cette **survigilance transitoire**, l'organisation doit accepter un coût supplémentaire de supervision humaine, supérieur à ce qui sera nécessaire en régime stabilisé. Sous-estimer cette période expose à des accidents de transition ; la prolonger indéfiniment installe une supervision rituelle où l'effet d'érosion décrit par Bainbridge s'installe sous couvert de prudence.

- ▷ **Les indicateurs de détection d'une dérive transitoire**. Comment l'organisation saura-t-elle que quelque chose ne va pas, avant qu'un accident ne le révèle ? Les indicateurs habituels — taux de fréquence, taux de gravité — sont trop lents et trop agrégés pour signaler une suraccidentalité transitoire localisée sur les processus assistés par l'IA.

  Il faut concevoir, dès le déploiement, des indicateurs spécifiques : taux de contestation des recommandations IA, fréquence des reprises de main manuelle, délai moyen entre alerte IA et validation humaine, taux d'événements survenus sur les processus IA-assistés comparé aux processus non assistés. Ces indicateurs ne sont pas un *reporting* supplémentaire : ce sont les capteurs de la trajectoire elle-même.

- ▷ **Les conditions de franchissement d'étape**. La trajectoire comporte des paliers — d'un cas d'usage à plusieurs, d'un niveau d'assistance plus assisté vers un niveau d'assistance plus délégué, d'un site pilote vers un déploiement généralisé. Qui décide du passage au palier suivant, sur quels critères, avec quelle réversibilité si l'étape franchie s'avère prématurée ? Le risque, en absence de gouvernance explicite de ces franchissements, n'est

pas que l'organisation aille trop vite — c'est qu'elle y aille sans le savoir, par accumulation de décisions locales dont personne n'a mesuré l'effet cumulé sur le système.

La voix la plus difficile à faire entendre pendant cette phase est celle qui dit *« il est encore trop tôt »* ou *« nous sommes allés trop loin »*. Plus les investissements sont engagés, plus les premiers résultats sont visibles, plus la pression organisationnelle pousse à avancer — et moins l'organisation est réceptive à un signal de ralentissement.

C'est précisément dans cette configuration que le dispositif de survigilance transitoire prend son sens : il donne une voix institutionnalisée à la prudence au moment où celle-ci est la plus difficile à exprimer spontanément. Le décideur qui conçoit ce dispositif à froid protège son organisation d'une tentation à laquelle lui-même pourrait céder à chaud.

Test de maturité

- ▷ Votre plan de déploiement IA distingue-t-il explicitement une phase de survigilance initiale — avec son périmètre, sa durée et ses ressources dédiées — d'un régime de croisière ?
- ▷ Avez-vous défini les indicateurs qui vous signaleraient une dérive transitoire avant qu'un incident ne la révèle ?
- ▷ Qui, dans votre organisation, a la légitimité formelle pour dire « ralentissons » ou « reprenons la main » sans conséquence sur sa trajectoire professionnelle ?
- ▷ Si vous deviez suspendre aujourd'hui le déploiement en cours sans dégrader la sécurité de l'installation, le pourriez-vous ?

7

Tensions et enjeux : mise en débat

Nous avons dans ce document délibérément évité deux postures : l'enthousiasme technologique qui voit dans l'IA la solution aux fragilités structurelles de la sécurité industrielle, et le refus conservateur qui en ferait un risque supplémentaire à éviter. La réalité est plus inconfortable : l'IA est à la fois un outil puissant et un révélateur exigeant. Elle peut renforcer ce qui fonctionne et fragiliser ce qui tient par l'habitude. Elle oblige à expliciter ce qui était implicite, à choisir ce qui était reporté et à assumer ce qui était délégué au consensus silencieux.

Ce dernier chapitre ne conclut pas. Il ouvre le débat sur cinq tensions que ce document a identifiées, mais qu'il ne peut pas résoudre seul — parce qu'elles engagent des choix de valeurs, pas seulement des choix techniques. Ces tensions sont formulées comme des dilemmes : dans chaque cas, les deux positions sont légitimes et le curseur optimal dépend du contexte, du secteur et des arbitrages que chaque organisation est prête à assumer. Ce sont des tensions à piloter — pas à résoudre.

7.1 Tension 1 : transparence ou performance ?

Les modèles d'IA les plus performants sont souvent les plus opaques. Les réseaux de neurones profonds et les grands modèles de langage atteignent des niveaux de précision remarquables sur la détection d'anomalies, l'analyse de textes et la reconnaissance de situations — mais personne ne peut expliquer complètement pourquoi ils produisent tel résultat plutôt que tel autre. Les modèles les plus explicables — réseaux bayésiens, systèmes à règles, arbres de décision — sont traçables et démontrables, mais moins puissants sur les données complexes et non structurées.

Pour les industries à haut risque, cette tension est structurelle. Le régulateur exige la traçabilité et la démontrabilité — ce qui oriente vers les modèles explicables. Mais les fragilités informationnelles décrites dans le chapitre 1 — textes non structurés, données hétérogènes, correspondance sémantique entre sources — nécessitent précisément les modèles opaques pour être traitées. Se limiter aux modèles explicables par prudence réglementaire, c'est renoncer aux apports les plus prometteurs. Déployer les modèles opaques sans cadre de démonstration, c'est fragiliser la confiance du régulateur et du public.

Mise en débat

Ce dilemme appelle un travail doctrinal entre exploitants et régulateurs. Que signifie « démontrer la sûreté » pour un système intrinsèquement non-déterministe ? Peut-on accepter une démonstration probabiliste « ce modèle se trompe dans moins de X% des cas » là où l'on exigeait une démonstration déterministe « ce système fait toujours ce pour quoi il a été conçu » ? Et si oui, pour quels cas d'usage et à quelles conditions ?

7.2 Tension 2 : assister ou affaiblir ?

Le paradoxe de Bainbridge traverse tout ce *Cahier* : l'IA qui aide en routine affaiblit l'humain pour l'exceptionnel. **Le blanchiment de légitimité** en est la forme organisationnelle — la signature humaine sur une décision effectivement algorithmique. Mais l'alternative — renoncer à l'assistance pour préserver les compétences — condamnerait les organisations à traiter manuellement une complexité informationnelle qu'elles ne maîtrisent déjà plus. Les six fragilités du chapitre 1 montrent que le statu quo n'est pas tenable non plus.

La tension n'est donc pas entre IA et pas d'IA. Elle est entre deux formes de vulnérabilité : la vulnérabilité par surcharge informationnelle, où l'humain ne peut plus traiter ce qu'il devrait traiter, et la vulnérabilité par atrophie, où l'humain ne sait plus faire ce qu'il devrait savoir faire.

Le curseur entre les deux n'est probablement pas le même selon les secteurs, les sites, les métiers et les situations. Le chapitre 3 a proposé le continuum outil/conseiller/producteur comme grille de positionnement. Mais le bon positionnement pour une installation donnée reste un choix local, contextuel et révisable.

Mise en débat

Existe-t-il un « seuil de dépendance » au-delà duquel l'assistance IA devient une vulnérabilité plutôt qu'un renfort ? Comment le mesurer ? Et qui, dans l'organisation, a la légitimité pour dire « nous sommes allés trop loin » — sachant que cette voix sera la plus difficile à faire entendre quand tout fonctionne bien ?

7.3 Tension 3 : partager ou protéger les données ?

Le chapitre 5 a montré que les données de l'entreprise étendue sont l'angle mort le plus critique. Mais partager les données entre donneur d'ordre et sous-traitant suppose une confiance qui n'existe pas encore — et qui pourrait être détruite par un usage asymétrique de l'IA. La cybersécurité ajoute une couche : chaque connexion créée pour alimenter un modèle élargit la surface d'attaque. La directive NIS2 étend les obligations de cybersécurité aux sous-traitants critiques. Le Data Act encadre le partage des données industrielles. Le devoir de vigilance questionne la responsabilité du donneur d'ordre sur l'ensemble de sa chaîne.

La tension est structurelle : la sécurité industrielle exige l'ouverture et le partage des données pour détecter les signaux faibles transversaux, pendant que la sûreté informatique et la protection des intérêts commerciaux exigent la fermeture et le cloisonnement. L'apprentissage fédéré et les référentiels anonymisés sont des pistes techniques. Mais la solution est d'abord relationnelle : un cadre de gouvernance partagé qui définit qui accède à quoi, pour quel usage, avec quelle réciprocité et sous quelle protection.

Mise en débat

Les industries à haut risque gagneraient-elles à coopérer entre elles — et avec leurs sous-traitants — pour partager les enseignements des incidents liés à l'IA, plutôt que de les traiter chacune dans le secret ?

Quand les incidents graves sont rares, mais critiques, le partage intersectoriel accélère l'apprentissage de tous. Mais quelle organisation est prête à reconnaître publiquement qu'un algorithme déployé dans son système de sécurité s'est trompé ?

7.4 Tension 4 : innover ou attendre ?

Les secteurs industriels qui retarderaient l'introduction de l'IA en attendant un cadre réglementaire stabilisé risquent de perdre en attractivité pour les spécialistes techniques dont les compétences sont déjà rares, et de prendre du retard sur des secteurs moins régulés qui capitaliseront l'expérience.

[Bieder et al. 2024] l'ont souligné : dans un contexte de déficit démographique et de perte d'attractivité du travail industriel, la capacité à intégrer les technologies nouvelles est un facteur de recrutement. Mais ceux qui avancent sans cadre prennent le risque de déployer des systèmes non qualifiables ou de créer des dépendances irréversibles à des technologies dont ils ne maîtrisent ni le cycle de vie ni les modes de défaillance.

Le chapitre 5 a proposé une voie médiane : commencer par les cas d'usage à faible risque réglementaire, en niveau d'assistance *outil* ou *conseiller*, pour construire l'expérience et la crédibilité avant d'aborder les cas critiques. Mais cette stratégie du « petit pas » a ses propres risques : les cas d'usage à faible risque sont aussi ceux à faible valeur ajoutée visible, ce qui peut décourager les directions et décrédibiliser la démarche avant qu'elle ait produit ses preuves sur les cas vraiment significatifs.

Mise en débat

Comment construire une stratégie d'adoption qui ne soit ni téméraire ni timorée ? Le dialogue précoce avec le régulateur est une piste — mais il suppose que le régulateur soit lui-même prêt à coconstruire un cadre sans le figer prématurément.

7.5 Tension 5 : l'IA change-t-elle la nature de la gestion de la sécurité industrielle ?

C'est la question la plus fondamentale et la moins tranchée. Les modèles de management de la sécurité développés depuis trente ans reposent sur une place prédominante de l'humain dans le pilotage du risque — c'est la partie « noble et intelligente » de la sécurité : le retour d'expérience, l'analyse du risque, les barrières de prévention, les démonstrations de sûreté.

La part de la technologie se limite aux machines, à leur conception, à leur fiabilité et à la qualité de leurs interfaces. Si l'IA transfère progressivement aux machines une partie du pilotage cognitif de la sécurité — le diagnostic, la détection, la recommandation, la simulation — alors les modèles FOH qui fondent le management de la sécurité doivent être repensés.

Ce *Cahier* a proposé une réponse partielle à cette question : la grille déléguer/assister/sanctuariser, le continuum outil/conseiller/producteur, et l'IA comme miroir du modèle de sécurité. Mais cette réponse est ancrée dans le monde d'aujourd'hui — où l'IA est un outil d'aide à la décision, pas un agent autonome. Si les évolutions en cours — modèles multimodaux, agents autonomes capables d'enchaîner des actions sans supervision intermédiaire, IA en temps réel — se concrétisent dans les industries à risque, la frontière entre « assister » et « déléguer » se déplacera, et la question de ce que l'on choisit de « sanctuariser » deviendra encore plus critique. Un agent qui optimise un processus de sécurité peut adopter des trajectoires que ses concepteurs n'avaient pas anticipées (la Région 4 du document CANUKUS — impact significatif et forte autonomie) appellera alors des barrières conventionnelles indépendantes à un niveau de rigueur que les cadres actuels n'exigent pas encore.

Une autre ligne de fond mérite d'être signalée. Les cadres qui ont structuré le management de la sécurité industrielle depuis trente ans ont reposé sur une identification claire de la sécurité industrielle comme fonction distincte — avec ses indicateurs, ses métiers, ses budgets, sa gouvernance propre. À mesure que l'IA traite simultanément la performance opérationnelle, la maintenance, la qualité et la sécurité à partir des mêmes gisements de données, la frontière entre ces registres se brouille.

La promesse — formulée par certains — est que la performance en sécurité finira par se fondre dans la performance globale, portée par des outils qui ne distinguent plus les finalités. Cette perspective peut être lue comme l'aboutissement du mouvement d'intégration qualité-sécurité-environnement amorcé depuis deux décennies. Elle peut aussi être lue comme la disparition d'une vigilance spécifique, celle qui consistait précisément à maintenir la sécurité comme préoccupation autonome et non négociable face aux pressions de performance.

Mise en débat

[Bieder et al. 2024] invitent à penser le management de la sécurité à l'ère du « vivre avec » l'incertitude et la complexité. Ce document a montré que l'IA est l'un des vecteurs de cette complexité nouvelle — et peut-être aussi l'un des outils pour y faire face. Deux questions restent ouvertes :

- ▷ Les industries à haut risque sauront-elles utiliser l'IA pour réduire les écarts entre leurs trois modèles de sécurité plutôt que pour en créer un quatrième — le modèle algorithmique — qui ajouterait une couche d'opacité à un système déjà complexe ?
- ▷ Et si la performance sécurité finit par se fondre dans la performance globale, qui portera la voix qui dit encore « ceci n'est pas négociable » quand toutes les métriques convergeront sur le même tableau de bord ?

8

Conclusion

Ce document a été écrit dans une posture **d'anticipation éclairée**. La meilleure façon de se préparer à l'IA n'est pas de courir après la technologie — c'est de faire le travail que l'IA oblige à faire : expliciter son modèle de sécurité, réduire les écarts entre ce qui est prévu, ce qui est compris et ce qui est fait, et préserver la capacité humaine de juger, de contester et de décider.

Une quatrième contribution, pas un quatrième pilier

Pris dans une perspective plus longue, ce qui se joue avec l'IA s'inscrit dans une trajectoire de plusieurs décennies. La sécurité industrielle a d'abord été pensée comme une affaire technique — maîtrise des équipements, ingénierie des barrières, conception des installations. Elle a ensuite intégré la dimension du système de management — formalisation des processus, évaluation des risques, pilotage par les indicateurs. Elle s'est enfin ouverte aux FOH — travail réel, compétences collectives, culture de signalement. Chacune de ces trois vagues a contribué, à son échelle, à faire reculer la fréquence des accidents graves et mortels, et chacune a buté sur ce que la précédente ne savait pas traiter.

L'IA n'est pas un quatrième pilier qui s'ajouterait aux trois précédentes. C'est un outil qui peut renforcer chacun d'entre eux : la maintenance prédictive sert la technique, la détection de signaux faibles sert le système de management, l'analyse du travail réel sert les FOH.

Sa promesse — plausible, mais non encore démontrée à l'échelle de la prévention des accidents graves — est de traiter la dimension informationnelle que les trois vagues précédentes atteignaient difficilement : données dispersées, signaux non corrélés, mémoire organisationnelle érodée. Si cette promesse se confirme, l'IA aura contribué à ce mouvement historique en démultipliant l'efficacité de ce qui existe déjà — pas en le remplaçant.

La prévention des accidents graves et majeurs précède l'IA

Une conclusion structurelle traverse les analyses de ce document et mérite d'être formulée explicitement, même si elle est inconfortable. Le déploiement de l'IA dans une organisation à haut risque ne produit ses bénéfices durables qu'à une condition : que l'organisation continue à entretenir ce qui produit la connaissance commune — présence préventeur dimensionnée, présence terrain managériale active, culture de sécurité robuste, mécanismes institutionnels d'agrégation. Sans ces conditions, l'IA accélère statiquement la qualité des décisions individuelles tout en érodant dynamiquement la matière même qui la rend pertinente.

Cette conclusion conduit à un paradoxe que nous ne prétendons pas résoudre. Les organisations qui auraient le plus à gagner d'un déploiement IA — celles dont les fragilités informationnelles sont les plus marquées — sont aussi celles où le déploiement est le plus dangereux. Inversement, les organisations qui peuvent absorber l'IA sans risque structurel sont celles dont les mécanismes humains traitent déjà partiellement leurs fragilités, et pour qui les bénéfices sont plus modestes.

L'implication stratégique est claire. La priorité d'une organisation à haut risque face à l'IA n'est pas le déploiement, c'est la consolidation préalable de la prévention des accidents graves et majeurs. Le travail décrit dans le Cahier *Donneur d'ordre, architecte de la culture de sécurité* — explicitation du modèle de sécurité, dimensionnement de la fonction prévention, présence terrain, culture juste, observatoire — n'est pas un programme parallèle à celui du déploiement

IA. C'est sa condition de possibilité. Et le diagnostic préalable proposé au chapitre 7 n'est pas un alourdissement administratif : c'est l'acte qui détermine si l'organisation est en état de bénéficier de l'IA ou si elle s'apprête à élargir ses propres brèches.

Cette articulation entre les deux Cahiers n'a pas été conçue à l'avance. Les fondamentaux qui font la robustesse d'une organisation à haut risque face à l'IA sont les mêmes que ceux qui font sa robustesse face aux accidents graves et mortels. Cela n'est pas un hasard — c'est probablement le résultat le plus important de ce *Cahier*. L'IA ne change pas la nature de la sécurité industrielle ; elle teste sa robustesse selon des modalités nouvelles, mais à travers les mêmes fondamentaux.

Point clé

Pour le praticien, cela signifie qu'il n'y a pas deux feuilles de route distinctes — l'une pour la culture de sécurité, l'autre pour l'IA. Il y a une seule trajectoire de transformation, qui consolide d'abord ce qui rend l'organisation capable de se voir elle-même, et qui déploie ensuite l'IA dans le périmètre que cette capacité rend soutenable. L'inverser, c'est risquer que l'IA aggrave silencieusement ce qu'elle prétendait soulager.

Un révélateur et une épreuve

Si la PAGEM précède l'IA, alors l'IA n'est pas un sujet à part — c'est un test appliqué aux fondamentaux. L'auteur ne prétend pas avoir toutes les réponses ; il a identifié des questions, parce qu'il les affronte depuis quarante ans sur le terrain. L'IA n'est ni *la* solution aux fragilités structurelles de la sécurité industrielle, ni une menace nouvelle à conjurer. C'est un révélateur et une épreuve.

- **Un révélateur**, parce qu'elle oblige à expliciter ce que les organisations laissent implicite depuis des années : leur modèle de sécurité, leurs arbitrages entre production et sûreté, leurs croyances sur ce qui les protège, les écarts entre ce qui est prévu, ce qui est compris et ce qui est fait. Ce travail d'explicitation a une vertu propre, indépendante de la technologie : il rend l'organisation plus sûre.
- **Une épreuve**, parce qu'elle teste la capacité de l'organisation à faire les bons choix — ce qu'elle confie à la machine, ce qu'elle garde sous contrôle humain, ce qu'elle protège de toute interférence algorithmique — et à maintenir, face à un outil puissant et séduisant, la capacité humaine de juger, de contester et de décider.

A

Glossaire

Termes liés à l'intelligence artificielle

Algorithme — Séquence d'instructions mathématiques qu'un programme informatique exécute pour résoudre un problème ou produire un résultat. En IA, l'algorithme « apprend » à partir des données plutôt que de suivre des règles écrites à l'avance.

Apprentissage fédéré (*federated learning*) — Technique permettant d'entraîner un modèle d'IA sur les données de plusieurs organisations sans que les données brutes ne quittent chaque système d'information. Chaque participant entraîne le modèle localement et partage les paramètres appris, pas les données.

Biais d'automation — Tendance cognitive à suivre les recommandations d'un système automatisé même lorsque d'autres indices suggèrent le contraire. Phénomène individuel, à distinguer du blanchiment de légitimité qui est un phénomène organisationnel.

Boîte noire — Système dont le fonctionnement interne n'est pas accessible ou compréhensible par l'utilisateur. Les réseaux de neurones profonds et les LLM sont des boîtes noires : ils produisent un résultat sans que l'on puisse expliquer complètement le raisonnement qui y conduit.

Data drift/Model drift — Dégradation progressive de la performance d'un modèle d'IA. Le *data drift* survient quand les données opérationnelles s'écartent des données sur lesquelles le modèle a été entraîné. Le *model drift* survient quand le modèle ne représente plus fidèlement les phénomènes sous-jacents. Les deux conduisent à une détérioration silencieuse — le système se dégrade sans le signaler.

Grand modèle de langage (*Large Langage Model*, LLM) — Modèle d'IA entraîné sur de vastes corpus de textes, capable de comprendre et de générer du langage naturel. Utilisé pour la recherche sémantique, l'analyse de textes, les assistants conversationnels. Peut produire des réponses plausibles, mais fausses (*hallucinations*).

Hallucination — Production par un modèle d'IA d'informations fausses ou inventées, présentées avec la même assurance que des informations correctes. Risque particulièrement critique dans les contextes de sécurité industrielle.

Historian (*data historian*) — Base de données spécialisée dans le stockage de séries temporelles à haute fréquence provenant des systèmes de supervision industrielle (SCADA). Contient l'historique des paramètres process (températures, pressions, débits).

Jumeau numérique — Réplique virtuelle d'une installation, d'un équipement ou d'un processus, alimentée en temps réel par les données opérationnelles. Permet de simuler des scénarios, de tester des modifications et d'anticiper des comportements sans intervenir sur l'installation réelle.

Machine learning (ML) — Sous-domaine de l'IA dans lequel les algorithmes apprennent des régularités à partir de données, sans être explicitement programmés pour chaque cas. Inclut l'apprentissage supervisé (classification, régression), non supervisé (clustering) et par renforcement.

NLP (*Natural Language Processing*) — Traitement automatique du langage naturel. Ensemble de techniques d'IA permettant à une machine de lire, comprendre et extraire du sens dans

des textes écrits en langage humain. Utilisé pour l'analyse des rapports de REX, la recherche sémantique, la détection de contradictions dans les référentiels.

RAG (*Retrieval-Augmented Generation*) — Architecture qui combine la recherche documentaire et la génération de texte par un LLM. Le système retrouve d'abord les documents pertinents dans une base, puis génère une réponse contextualisée à partir de ces documents. Réduit le risque d'hallucination en ancrant la réponse dans des sources identifiées.

Cotation— Attribution automatique d'un score numérique à une situation, un risque, un niveau de risque à maîtriser, sur la base de critères pondérés par un algorithme. La cotation n'est pas une décision — c'est un signal qui doit être interprété par un humain.

Vision par ordinateur — Famille de techniques d'IA permettant à une machine d'analyser des images ou des flux vidéo pour détecter des objets, reconnaître des situations ou mesurer des écarts. Applications en sécurité : détection du port d'EPI, contrôle de conformité de gestes, surveillance de zones à risque.

Termes liés à la sécurité industrielle

Alarm flood — Situation de saturation dans laquelle un opérateur reçoit un volume d'alarmes qu'il ne peut plus trier ni traiter. Documenté comme facteur contributif dans de nombreux accidents industriels. Normes de référence : EEMUA 191, ISA-18.2.

Barrière de défense — Dispositif technique, organisationnel ou humain destiné à prévenir un événement indésirable ou à en limiter les conséquences. La défense en profondeur repose sur l'empilement de barrières indépendantes.

Culture juste — Approche dans laquelle l'organisation distingue l'erreur (non sanctionnée, analysée pour apprendre) de la violation délibérée (sanctionnée). Condition nécessaire à la remontée des signaux faibles et au retour d'expérience.

Défense en profondeur — Principe de conception selon lequel la sécurité repose sur plusieurs niveaux de protection indépendants, de sorte qu'aucune défaillance unique ne puisse conduire à un accident.

Droit de STOP — Droit reconnu à tout intervenant d'arrêter une activité s'il estime qu'une condition de sécurité n'est pas remplie, sans conséquence disciplinaire et avec l'obligation d'analyser la situation avant toute reprise. Transposé dans ce document au domaine de l'IA : le doute bénéficie au contestataire, pas à l'algorithme.

EPI — Équipement de protection individuelle (casque, lunettes, gants, harnais, etc.).

HIPO/SHPG — Événement à haut potentiel de gravité (*High Potential Incident*/*Situation à Haut Potentiel de Gravité*). Événement qui aurait pu avoir des conséquences graves dans des circonstances légèrement différentes, même s'il n'a pas produit de dommage.

LOTO (*Lock Out/Tag Out*) — Procédure de consignation qui garantit qu'un équipement est isolé de ses sources d'énergie et ne peut pas être remis en service pendant une intervention. Séquence critique pour la sécurité dont la conformité peut être vérifiée par vision par ordinateur.

MOC (*Management of Change*) — Processus formalisé de gestion des modifications d'une installation, d'un procédé ou d'une organisation. Inclut une analyse d'impact sur la sécurité avant toute mise en œuvre.

Normalisation de la déviance — Processus par lequel un écart par rapport à la norme, initialement perçu comme anormal, est progressivement accepté comme normal par l'organisation parce qu'il n'a pas produit de conséquence visible. Concept issu de l'analyse de l'accident de la navette Challenger **[Vaughan 1996]**.

PAGEM — Prévention des Accidents Graves Et Mortels. Cadre développé par l'Icsi structurant les activités de sécurité en six domaines : préparation, retour d'expérience, situations à haut potentiel de gravité, règles qui sauvent, marges de manœuvre, audits et terrain.

REX — Retour d'expérience. Processus systématique de collecte, d'analyse et de partage des enseignements tirés des événements (accidents, incidents, quasi-accidents) pour prévenir leur récurrence.

RQS — Règles qui sauvent. Ensemble restreint de règles dont le non-respect peut directement conduire à un accident grave ou mortel. Leur transgression appelle une réponse managériale immédiate.

Signal faible — Information fragmentaire ou ambiguë qui, prise isolément, ne déclenche pas d'alerte, mais qui, rapprochée d'autres signaux, peut révéler une dérive ou un risque émergent.

Termes liés aux systèmes d'information industriels

ERP (*Enterprise Resource Planning*) — Progiciel de gestion intégré couvrant les processus de l'entreprise (production, achats, RH, finance). Contient les données de planification et de coactivité.

GED — Gestion électronique des documents. Système de stockage, de classement et de recherche des documents (procédures, modes opératoires, consignes).

GMAO/CMMS — Gestion de maintenance assistée par ordinateur (*Computerized Maintenance Management System*). Système contenant l'historique des interventions de maintenance, les fiches d'équipement, les ordres de travail.

LMS (*Learning Management System*) — Plateforme de gestion de la formation. Contient les historiques de formations suivies, les résultats d'évaluations, les taux de couverture des formations réglementaires.

SCADA/DCS (*Supervisory Control and Data Acquisition / Distributed Control System*) — Système de supervision et de contrôle-commande d'une installation industrielle. Collecte en temps réel les données des capteurs (températures, pressions, débits, états d'équipements).

SI HSE/SSE — Système d'information Hygiène, Sécurité, Environnement (ou Santé, Sécurité, Environnement). Base de données contenant les événements de sécurité, les analyses d'incidents, les plans de prévention, les observations terrain.

SIRH — Système d'information des ressources humaines. Contient les habilitations, les certifications, les matrices de compétences, les plannings.

SIS (*Safety Instrumented System*) — Système instrumenté de sécurité. Dispositif automatique de protection qui agit indépendamment du système de contrôle-commande pour mettre l'installation en état sûr en cas de dépassement de seuils critiques.

Concepts propres à ce document

Blanchiment de légitimité — Pathologie organisationnelle dans laquelle la signature humaine sur une décision effectivement produite par l'IA donne une apparence de gouvernance sans la substance. Se distingue du biais d'automation — phénomène cognitif individuel **[Parasuraman et Manzey 2010]** — et de la supervision rituelle — glissement procédural individuel pointé par **[EDPS 2025]**. Le blanchiment de légitimité est le résultat agrégé de ces deux mécanismes à l'échelle d'une institution qui perd sa capacité à assigner la responsabilité. Les trois appellent des parades différentes : formation individuelle, conception des processus, gouvernance et espace de contestation formalisé.

Continuum outil/conseiller/producteur — Trois niveau d'assistance au sein de la catégorie « assister » de la grille DAS. Dans le niveau *outil*, l'humain pilote l'IA. Dans le niveau *conseiller*, l'IA propose et l'humain décide. Dans le niveau *producteur*, l'IA produit et l'humain supervise. Plus on avance sur ce continuum, plus les impacts FOH s'aggravent et plus le risque de blanchiment de légitimité augmente.

Déléguer/assister/sanctuariser — Grille de décision inspirée de **[Hazan et Sibony 2026]** et de **[Parasuraman et al. 2000]**, structurant les choix de déploiement de l'IA en trois catégories : ce que l'on confie à la machine (déléguer), ce que l'on fait avec son aide (assister), et ce que l'on protège de toute interférence algorithmique (sanctuariser).

Espace de contestation — Dispositif organisationnel dans lequel des acteurs désignés ont le droit et le devoir de contester un résultat produit par l'IA, avec la protection institutionnelle nécessaire pour exercer ce droit sans conséquence. Transposition au domaine de l'IA du droit de STOP.

IA comme miroir — Thèse centrale du présent document : l'IA, en traitant les données de l'organisation, révèle les écarts entre les trois modèles de sécurité (prescrit, managérial porté, opérant) et rend visibles des croyances implicites, des dérives normalisées et des angles morts que l'organisation n'avait pas quantifiés.

Modèle prescrit/modèle managérial porté/modèle opérant — Trois modèles de sécurité qui coexistent dans toute organisation. Le modèle *prescrit* est ce qui est écrit (procédures, consignes, référentiels). Le modèle *managérial porté* est ce que les manager croient et porte. Le modèle *opérant* est ce qui se fait réellement sur le terrain. Les écarts entre ces trois modèles sont une source majeure de vulnérabilité. Réduire les écarts entre ces modèles est le cœur de la PAGEM.

Supervision rituelle — Expression de **[EDPS 2025]** désignant une supervision humaine qui a perdu sa substance : l'opérateur clique « approuver » parce que le système le recommande, sans exercer réellement son jugement. Forme institutionnalisée du blanchiment de légitimité.

Acronymes réglementaires et institutionnels

AIEA/IAEA — Agence internationale de l'énergie atomique.

AI Act — Règlement européen (UE) 2024/1689 établissant des règles harmonisées en matière d'intelligence artificielle. Application complète août 2026.

CANUKUS — Acronyme désignant la coopération trilatérale entre les régulateurs nucléaires du Canada (CNSC), du Royaume-Uni (ONR) et des États-Unis (NRC).

CNSC — Canadian Nuclear Safety Commission (Commission canadienne de sûreté nucléaire).

COFSOH — Comité d'orientation Facteurs Organisationnels et Humains/Sûreté.

Data Act — Règlement européen (UE) 2023/2854 sur l'accès équitable aux données et leur utilisation.

EDPS — European Data Protection Supervisor (Contrôleur européen de la protection des données).

FOH — Facteurs Organisationnels et Humains.

HRO — *High Reliability Organization.* Organisation à haute fiabilité.

NIS2 — Directive européenne (UE) 2022/2555 sur la sécurité des réseaux et des systèmes d'information.

NRC — US Nuclear Regulatory Commission.

ONR — UK Office for Nuclear Regulation.

B

Méthodologie d'application de la grille DAS

La grille DAS, telle qu'elle est posée au § 3.2, est une grille de **délibération**, pas un système de classement. Des éléments structurent l'arbitrage : cinq critères tâche (gravité potentielle, révocabilité de l'erreur, régularité structurelle, barrières aval, nature de la décision). Les tests de viabilité (blanchiment de légitimité, effet Goodhart, ironie de Bainbridge) permettent ensuite de vérifier que le régime (Déléguer, assister, sanctuariser) et le niveau d'assistance (outil, conseiller, producteur) choisi tiennent dans la durée.

Cette annexe outille l'application opérationnelle de la grille pour les organisations qui souhaitent l'expérimenter sur leurs cas réels. Elle ne se substitue pas au chapitre 3, qui en énonce la doctrine ; elle propose une procédure d'instruction, un format de traçabilité, et une illustration résolue. Comme indiqué au § 3.2.2, la grille est une proposition de l'auteur qui a besoin d'être éprouvée par l'usage. Cette annexe est donc, elle aussi, un objet provisoire — elle évoluera à mesure que les retours d'expérience des organisations qui l'auront mobilisée seront partagés.

B.1 Procédure d'instruction d'un cas

L'application de la grille à un cas concret se conduit en six étapes. Chacune appelle une **justification** et non un score : c'est la trace du raisonnement qui structure la décision et qui rendra possible une revue ultérieure.

Étape 1 — Désagréger la fonction étudiée. Une fonction nommée *« maintenance prédictive »*, *« analyse REX »* ou *« vision par ordinateur sur chantier »* recouvre généralement plusieurs sous-fonctions dont le régime DAS peut différer. La granularité est elle-même un acte d'arbitrage : trop fine, elle multiplie les analyses sans valeur ajoutée ; trop grossière, elle masque les régimes (DAS) et niveau d'assistances (outil, conseiller, producteur) différenciés et conduit à un arbitrage moyen qui n'est juste pour aucune sous-fonction. La règle pratique : descendre au niveau où le profil sur les cinq critères tâche devient homogène.

Étape 2 — Renseigner les cinq critères tâche pour chaque sous-fonction identifiée. L'évaluation se formule en *favorable/défavorable/mitigé*, accompagnée d'une justification courte qui explicite ce qui fonde l'appréciation. La justification est ce qui distingue la grille d'une checklist.

Étape 3 — Appliquer la règle de priorité absolue. Si le critère 5 (nature de la décision) est défavorable — c'est-à-dire si la décision engage un jugement de valeur, une responsabilité ultime ou un arbitrage entre objectifs incommensurables — alors la sous-fonction est sanctuarisée, indépendamment du profil sur les autres critères. L'instruction s'arrête là pour cette sous-fonction.

Étape 4 — Choisir le régime et sa profondeur. Pour les sous-fonctions non sanctuarisées par l'étape 3, le profil agrégé des critères 1 à 4 désigne le régime (déléguer vs assister) et, si assister est retenu, le niveau d'assistance (outil, conseiller, producteur). Les conditions du niveau d'assistance *producteur* — tâche fortement standardisée, erreur révocable, dispositif de validation humaine effective — doivent être explicitement satisfaites.

Étape 5 — Tracer le raisonnement. Le résultat de l'instruction est consigné dans un document daté, signé, et révisable. Ce document précise pour chaque sous-fonction le régime et le niveau d'assistance retenus, les justifications associées, et les conditions de revue ultérieure (périodicité, événements déclencheurs, indicateurs à surveiller). C'est ce document qui rend possible l'exercice ultérieur des tests de viabilité.

B.2 Procédure de revue périodique des trois tests de viabilité

Un arbitrage DAS juste à l'instant t peut dériver dans la durée. La revue périodique structurée des trois tests de viabilité est ce qui rend la grille **opérante** dans le temps long. Elle est aussi importante que l'instruction initiale.

Périodicité recommandée : annuelle, complétée par une revue *ad hoc* en cas d'événement déclencheur (incident, alerte sur la dérive d'un indicateur, changement de fournisseur ou de version du système IA, départ massif des compétences humaines associées).

Acteurs à convoquer : le porteur opérationnel du dispositif IA, un représentant de la fonction sécurité, un représentant du métier concerné, et — c'est essentiel — au moins un acteur désigné institutionnellement comme *contesteur*, c'est-à-dire mandaté pour dire *« ce que je vois sur le terrain ne correspond pas à ce que dit le système »* sans conséquence hiérarchique.

Questions opératoires pour chaque test :

Test	Question opératoire pour la revue
Blanchiment de légitimité	Quel est le taux de non-validation humaine ? Les non-validations sont-elles motivées et traçables ? Existe-t-il des cas documentés où un humain a contesté avec succès une sortie de l'IA ?
Effet Goodhart	L'indicateur de performance de l'IA progresse-t-il pendant que les événements terrain corrélés progressent ou régressent ? Le proxy reste-t-il fidèle à l'objectif réel de sécurité ?
Ironie de Bainbridge	Les compétences humaines préservées par la sanctuarisation sont-elles exercées suffisamment fréquemment ? Existe-t-il un plan d'entretien (exercices hors assistance, simulations) et est-il tenu ?

Décisions possibles à l'issue de la revue : statu quo, ajustement de la profondeur dans le niveau d'assistance (passage outil ↔ conseiller ↔ producteur), bascule de régime (assister ⟶ sanctuariser, ou délégation ⟶ assister), arrêt du dispositif si la viabilité n'est plus tenue. Toute décision de revue est tracée selon le même format que l'instruction initiale, et le document est versionné.

B.3 Limites de la grille et angles morts

Cette annexe propose un outillage opérationnel d'une grille qui reste, par construction, un outil de délibération. Six limites doivent être nommées :

- La grille traite des **systèmes IA système porteur d'un modèle de sécurité**, au sens du chapitre 2. Elle ne traite pas des IA *outil* (capteurs intelligents…) dont la gouvernance relève de logiques propres.
- La grille ne traite pas des IA t-elle que robotique d'inspection, automatismes de protection déterministes, télé-opération).
- La grille **suppose que l'organisation a explicité son modèle de sécurité**. Sans cette explicitation, l'arbitrage DAS s'appuie sur des prémisses implicites qui rendent la traçabilité fictive.
- La grille **est un outil de délibération, pas de classement automatique**. Son usage suppose un collectif compétent capable de la mobiliser. Une organisation qui réduirait la grille à un tableau à cocher reproduirait exactement l'effet Goodhart qu'elle dénonce.

Enfin, la grille n'épuise pas les conditions de réussite d'un déploiement IA en sécurité industrielle. Les questions de cybersécurité, de souveraineté des données, de qualification réglementaire (§ 3.3) et de gouvernance d'ensemble du déploiement (chapitre 6) requièrent des instructions complémentaires que la grille ne porte pas.

C

Cartographie des familles d'IA et des sources de données

Cette annexe réunit deux cartographies complémentaires. La première recense les familles technologiques d'IA mobilisables en industrie à risque, avec leur maturité et leur contribution typique à la sécurité. La seconde recense les sources de données disponibles dans les systèmes d'information d'un site industriel, organisées par domaine PAGEM. Les deux cartographies se lisent indépendamment, mais se complètent : les familles d'IA décrites en partie 1 s'appliquent aux sources de données décrites en partie 2 pour servir les finalités informationnelles identifiées au chapitre 2.

Les sept familles d'IA informationnelles

Ce premier tableau résume les principales familles technologiques d'IA qui produisent de l'information et outillent la décision.

Famille d'IA	Usage principal	Maturité	Apport sécurité	Risque principal
ML supervisé/non supervisé	Maintenance prédictive, détection d'anomalies	★★★	Indirect	Dérive des données, cygnes noirs
Deep Learning	Diagnostic, prédiction complexe	★★☆	Indirect	Boîte noire, non démontrable
Vision par ordinateur	Surveillance EPI, zones, comportements	★★★	Direct	Surveillance vs sécurité
NLP/LLM	Analyse REX, extraction causes, conformité	★★☆	Direct	Hallucinations, atrophie du jugement
Jumeaux numériques	Simulation, formation, optimisation	★☆☆	Indirect	Fausse assurance, cybersécurité
IA symbolique/Bayésien	Modélisation des risques, EPS	★★★	Indirect	Rigidité, coût d'acquisition
VR/AR pilotée par IA	Formation aux situations dégradées, aide embarquée	★★☆	Indirect	Risque d'illusion de maîtrise

Une famille complémentaire — la robotique autonome ou semi-autonome — soustrait physiquement l'humain au risque (drones d'inspection, robots en zone irradiée, télé-opération). Elle relève d'un autre périmètre, celui des IA qui agissent matériellement à la place de l'intervenant, et n'est pas développée dans ce *Cahier*.

Les sources de données

Cette seconde cartographie recense les principales sources de données disponibles dans les systèmes d'information d'un site industriel à risque, organisées par domaine PAGEM. Elle complète les dimensions informationnelles décrites dans les fragilités du chapitre 1 et fournit au lecteur une vue concrète des gisements de données exploitables pour chaque cas d'usage IA.

Les données sont organisées en trois couches, qui correspondent à la réalité des architectures SI des industriels : les systèmes d'information de gestion (GMAO, ERP, SI HSE, SIRH, LMS, GED — données structurées et semi-structurées), les systèmes de supervision terrain (SCADA, historians, capteurs IoT, caméras, alarm management — données temps réel, souvent en séries temporelles), et les données croisées dont la valeur analytique provient du rapprochement entre les deux couches précédentes — un croisement rarement automatisé dans les entreprises.

Avertissement : Les données de sécurité ne sont pas neutres. Elles sont produites par une culture, un système de remontée, des asymétries hiérarchiques, une économie du silence. Chaque source décrite ci-dessous porte les biais du système qui l'a produite.

Trois constats structurants

Les données existent, mais elles sont éparpillées. Pour chacun des six domaines, les entreprises à risque disposent de multiples sources de données pertinentes. Le problème n'est généralement pas l'absence de données, mais leur cloisonnement : chaque système a sa propre logique, ses propres référentiels, ses propres temporalités. Le croisement — qui est la source de valeur analytique maximale — est rarement automatisé.

Les données les plus riches pour la sécurité sont les moins structurées. Les descriptions libres des REX, les commentaires de techniciens dans les ordres de travail, les cahiers de quart, les comptes rendus d'entretiens FOH — c'est là que se trouvent les informations les plus riches sur le « pourquoi » des événements. Or ce sont ces données non structurées qui nécessitent le NLP pour être exploitées à échelle — et ce sont celles qui portent le plus fortement les biais de la culture de remontée.

Les données de l'entreprise étendue restent un angle mort. La sous-traitance représente souvent 30 à 60% des heures travaillées sur les sites à risque. Or les données des prestataires restent largement cloisonnées dans leurs propres systèmes d'information, échangées en PDF, sans structuration permettant un traitement analytique. C'est un chantier entier à ouvrir, traité dans le chapitre consacré à l'entreprise étendue.

Sources de données pour les six domaines de la prévention des accidents graves ou mortels

Domaine	SI de gestion	Supervision terrain	Exemples de croisements à valeur ajoutée
Préparation	◆ SI HSE (REX, plans de prévention), ◆ GMAO (historique interventions), ◆ SIRH (habilitations, compétences), ◆ LMS (formations, scores), ◆ Permis de travail, GED (procédures)	◆ SCADA/Historian (état équipements, conditions opératoires), ◆ Capteurs météo, ◆ Géolocalisation/contrôle d'accès, ◆ SIG (réseaux)	◆ REX × Conditions opératoires (cotation du niveau de risque à maîtriser contextualisé). ◆ Habilitations × Complexité intervention (adéquation compétences). ◆ Permis × REX co-activité (alertes dès l'autorisation)
REX	◆ SI HSE (base REX, descriptions libres, analyses causales, actions), ◆ GMAO (OT avec commentaires techniciens), ◆ GED (rapports d'enquête)	◆ SCADA/Historian (conditions au moment de l'événement) ◆ Vidéosurveillance (si disponible) ◆ Alarm management (alarmes avant l'événement)	◆ REX textuels × Données process (reconstruction du contexte) ◆ REX × SIRH (profils des équipes impliquées) ◆ Croisement inter-sites (détection de récurrences)
HIPO/SHPG	◆ SI HSE (base HIPO, barrières défaillantes, distance à l'accident), GMAO (défaillances ◆ équipements critiques, reports maintenance), ◆ Base EPS/EDD (scénarios, probabilités)	◆ SCADA (sollicitations SIS, dépassements de seuils) ◆ Capteurs gaz/feu ◆ Alarm management (alarm floods, alarmes chroniques) ◆ Données fatigue/charge de travail	◆ Sollicitations SIS × Scénarios EDD (combien de barrières restait-il ?) ◆ Reports maintenance × Barrières (indisponibilité coïncidente) ◆ Alarm floods × HIPO (surcharge cognitive comme précurseur)
Règles/RQS	◆ GED (référentiel de règles, révisions) ◆ SI HSE (écarts documentés, transgressions) ◆ SIRH/LMS (formations aux règles, scores) ◆ MOC (modifications d'installation)	◆ Vision IA (conformité EPI, postures, zones…), ◆ SCADA (séquences opératoires réelles vs prescrites), ◆ Capteurs LOTO (conformité consignation)	◆ Écarts (HSE) × Règles les plus transgressées (priorisation révision). ◆ Vision IA × REX (règles inapplicables vs inapliquées). ◆ Formations (LMS) × Écarts (efficacité réelle de la formation)
Marges de manœuvre	◆ SI HSE (signalements de situations où la règle est inapplicable), ◆ GMAO (dérogations, adaptations documentées), ◆ GED (REX positifs, adaptations réussies)	◆ SCADA (variabilité de la performance normale), ◆ Historian (dérives lentes de paramètres), ◆ Données de charge/planning (pression temporelle)	◆ Adaptations terrain × Résultats sécurité (distinguer écarts créateurs vs destructeurs). ◆ Pression temporelle × Événements (érosion des marges sous contrainte)
Audits et terrain	◆ SI HSE (rapports d'audit, observations terrain, indicateurs), ◆ GMAO (conformité maintenance), ◆ SIRH (couverture formations), ◆ ERP (indicateurs de performance)	◆ Vision IA (observations automatisées), ◆ Capteurs environnementaux (bruit, qualité air), ◆ Wearables (si déployés)	◆ Observations terrain × Indicateurs (cohérence entre ce qui est vu et ce qui est mesuré). ◆ Audits × REX (les constats d'audit prédisent-ils les événements ?)

Niveau de disponibilité typique par domaine

Ce tableau résume le niveau de disponibilité typique des données dans les entreprises à risque européennes de taille intermédiaire à grande.

Légende : ★★★ = disponible et couramment exploité ; ★★☆ = disponible, mais sous-exploité ; ★☆☆ = rarement disponible ou très cloisonné.

Domaine	SI HSE	GMAO	SCADA / Historian	Vision/IoT	Données croisées
Préparation	★★★	★★★	★★☆	★☆☆	★☆☆
REX	★★★	★★☆	★★☆	★☆☆	★☆☆
HIPO/SHPG	★★☆	★★★	★★★	★★☆	★☆☆
Règles/RQS	★★★	★☆☆	★★☆	★★★	★★☆
Marges	★☆☆	★☆☆	★★☆	★☆☆	★☆☆
Audits et terrain	★★★	★★☆	★☆☆	★★☆	★★☆

D

Synthèse du document trilatéral CANUKUS

Considerations for Developing Artificial Intelligence Systems in Nuclear Applications, Canadian Nuclear Safety Commission (CNSC), UK Office for Nuclear Regulation (ONR), US Nuclear Regulatory Commission (NRC) — Septembre 2024

Ce document est la première publication conjointe de régulateurs nucléaires internationaux sur l'intelligence artificielle. Il pose des principes — pas des exigences réglementaires — que l'ensemble des acteurs de la chaîne (développeurs, exploitants, intégrateurs, régulateurs) devrait considérer lors du déploiement de l'IA dans les applications nucléaires.

Cette synthèse en extrait les points clés pour le praticien de la sécurité industrielle et les met en correspondance avec les concepts développés dans ce document.

D.1 La grille à quatre régions : conséquences × autonomie

Le cœur du document est un modèle de catégorisation qui croise deux axes : la **signification d'une défaillance de l'IA** (impact minimal à impact significatif sur la sûreté) et le **degré d'autonomie accordé à l'IA** (de l'aide à la décision à l'autonomie complète). Ce croisement définit quatre régions.

Région	Caractéristiques	Exemples d'usage
Région 1 Faible impact, faible autonomie	L'IA informe la décision humaine. La défaillance a un impact minimal. Plus grande flexibilité de déploiement.	Analyse de données de maintenance, recherche documentaire, classification d'événements.
Région 2 Faible impact, forte autonomie	L'IA opère avec peu de supervision humaine. L'impact d'une défaillance reste limité, mais le temps de réaction humaine est réduit.	Optimisation opérationnelle, pilotage de paramètres non critiques.
Région 3 Impact significatif, faible autonomie	L'IA aide à la décision sur des fonctions critiques. La sortie peut être vérifiée avant action. Un processus de vérification robuste est requis.	Aide à la conception ou à la maintenance de systèmes de sûreté. Cotation de risque sur équipements classés.
Région 4 Impact significatif, forte autonomie	L'IA opère avec peu de temps pour la vérification humaine. Exige que d'autres composants robustes puissent atténuer une défaillance de l'IA.	Algorithmes de contrôle-commande optimisés par IA, protection automatisée.

Cette grille à quatre régions est fonctionnellement la même logique que la grille déléguer/assister/sanctuariser développée dans le chapitre 3 du Cahier, formulée dans le langage du régulateur. La région 1 correspond au régime « déléguer » sur des cas non critiques. La région 3 correspond au régime « assister » sur des équipements critiques. La région 4 est le domaine où la sanctuarisation et les barrières conventionnelles indépendantes sont nécessaires. Le continuum outil/conseiller/producteur affine la transition entre les régions.

D.2 Le principe de simplicité et les limites de la qualification

Le document pose un principe fort : la technologie la plus appropriée devrait être **aussi simple que possible** pour que les modes de défaillance puissent être analysés et compris, avec aussi peu de modes de défaillance inconnus que possible. En d'autres termes, l'IA ne doit pas compliquer inutilement un système qui fonctionnerait avec une technologie conventionnelle. Et lorsque l'IA est déployée, les principes éprouvés de la défense en profondeur — diversité, redondance, séparation, ségrégation — restent applicables : la sûreté ne doit jamais reposer entièrement sur un seul élément, qu'il soit humain ou algorithmique.

Point clé

Le document reconnaît explicitement qu'il n'existe actuellement aucune méthode pour quantifier la probabilité de défaillance d'un composant IA au sein d'un système. Cela rend difficile son intégration dans les études probabilistes de sûreté (EPS).

La fiabilité doit donc être dérivée de l'architecture du système global — c'est-à-dire de la capacité des composants conventionnels à compenser une défaillance de l'IA — plutôt que démontrée par la qualification du composant IA lui-même. C'est le paradoxe déterministe/non-déterministe identifié dans le chapitre 3.

D.3 Facteurs humains et organisationnels

Le document traite les FOH avec une profondeur remarquable pour un texte de régulateur. Quatre points méritent d'être soulignés :

- **Le continuum de confiance humain-machine** : Le document souligne qu'il faut établir un *niveau optimal de confiance* entre l'humain et l'IA — ni trop faible (l'humain ignore l'IA et perd son bénéfice) ni trop élevé (l'humain suit aveuglément l'IA et cesse d'exercer son jugement). L'excès de confiance conduit à la complaisance où l'humain ne surveille plus le système parce qu'il fonctionne habituellement bien — et n'est donc plus une sauvegarde efficace. C'est exactement la tension 2 (vigilance) de la grille FOH du *Cahier*.
- **L'opacité des modèles** : Le document note que les systèmes IA fonctionnent comme des « boîtes noires », à la différence des systèmes conventionnels où l'opérateur peut mémoriser les circuits logiques et identifier facilement un fonctionnement anormal. Cette opacité complique le sensemaking collectif en situation de crise — la tension 3 du *Cahier*.
- **Le maintien des qualifications** : Le document recommande de maintenir un *niveau minimal de qualification du personnel* pour les métiers augmentés par l'IA et de préserver une redondance humaine appropriée. C'est la parade à l'atrophie des compétences (tension 1 du *Cahier*) : l'enjeu est de préserver délibérément les compétences que l'IA rend moins nécessaires en routine.
- **L'impact sur la culture de sécurité** : Le document pose la question : comment vérifier, une fois le système déployé, que les décisions prises par l'IA sont cohérentes avec les priorités de sûreté de l'organisation ? Et comment les constats issus de la surveillance de la culture de sécurité seront-ils réinjectés dans la conception du système ? C'est la tension 5 du *Cahier* — et c'est aussi la question de l'espace de contestation développée dans le chapitre 6.

Lien avec le Cahier : Les quatre points FOH du document CANUKUS convergent avec les tensions de la grille FOH du Cahier. Cette convergence n'est pas fortuite : elle reflète le fait que les problèmes FOH posés par l'IA sont intrinsèques à la technologie, indépendants du secteur et du régulateur. Ce qui diffère, c'est la manière d'y répondre — et c'est précisément ce que nous proposons avec le continuum outil/conseiller/producteur et le concept de blanchiment de légitimité, qui ne figurent pas dans le document CANUKUS.

D.4 Cycle de vie, dérive et sécurité informatique

Le document insiste sur le fait que le cycle de vie d'un système IA est fondamentalement différent de celui d'un logiciel conventionnel. Deux phénomènes spécifiques à l'IA nécessitent une gestion continue : **la dérive des data** (les données opérationnelles s'écartent des données d'entraînement) et le **model dérive** (le modèle ne représente plus fidèlement les phénomènes sous-jacents). Les deux conduisent à une dégradation silencieuse de la performance — le système se détériore sans que rien le signale, sauf si des mécanismes de surveillance spécifiques sont en place.

Le document souligne également que le réentraînement d'un modèle IA comporte ses propres risques — surapprentissage, introduction de nouveaux biais — et que chaque mise à jour crée potentiellement un nouveau système qui devrait être requalifié. C'est un défi spécifique aux industries à haut risque, où le rythme d'évolution de l'IA (mises à jour fréquentes, amélioration continue) est en tension structurelle avec le rythme de la qualification de sûreté (rigoureuse, lente, documentée).

Sur la sécurité informatique, le document rappelle que les systèmes IA contiennent des composants provenant de sources externes ou ouvertes (données, logiciels, configurations matérielles), ce qui élargit la surface d'attaque. Des lignes directrices pour le développement sécurisé de systèmes IA ont été publiées conjointement par 22 agences de sécurité des trois pays et de leurs partenaires.

Point clé

Le document conclut que les normes consensus spécifiques à l'IA pour le domaine nucléaire ne seront probablement pas disponibles dans un avenir proche. En attendant, les normes nucléaires existantes restent un point de départ, complétées par la prise en compte des attributs spécifiques de l'IA. C'est précisément pourquoi le chapitre 6 du présent document recommande aux Codir d'engager le dialogue avec leur régulateur avant de déployer — pour coconstruire les critères plutôt que de les subir.

Bibliographie

Acemoglu, D., Kong, D. et Ozdaglar, A. (2026). *AI, human cognition and knowledge collapse*. Rapport technique, NBER. Working Paper 34910. **DOI** : 10.3386/w34910.

Amalberti, R. (1996). *La conduite de systèmes à risques*. Coll. Le Travail Humain. PUF, 2 édition. **ISBN** : 978-2130522775, 239 pages.

Amalberti, R. (2013). *Piloter la sécurité — Théories et pratiques sur les compromis et les arbitrages nécessaires*. Springer Verlag. **ISBN** : 978-2817803692, 145 pages.

Argyris, C. et Schön, D. A. (1978). *Organizational learning: a theory of action perspective*. Addison Wesley. **ISBN** : 978-0201001747, 356 pages.

Bainbridge, L. (1983). *Ironies of automation*. Automatica, 19(6):775–779. **DOI** : 10.1016/0005-1098(83)90046-8.

Bengio, Y., Lecun, Y. et Hinton, G. (2021). *Deep learning for AI*. Communications of the ACM, 64(7):58–65. **DOI** : 10.1145/3448250.

Bieder, C., Amalberti, R., Pariès, J. *et al.* (2024). *La sécurité à l'ère du « vivre avec » : Incertitude, complexité et nouvelles attentes*. Cahier de la sécurité industrielle 2024-05, Fondation pour une culture de sécurité industrielle. www.foncsi.org, **DOI** : 10.57071/420yzp.

Carlini, N., Tramèr, F., Wallace, E. *et al.* (2021). *Extracting training data from large language models*. Dans *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Carroll, J. S. (1998). *Organizational learning activities in high-hazard industries: the logics underlying self-analysis*. Journal of Management Studies, 35(6):699–717. **DOI** : 10.1111/1467-6486.00116.

Davies, D. (2025). *The Unaccountability Machine: Why Big Systems Make Terrible Decisions—and How the World Lost Its Mind*. The University of Chicago Press. **ISBN** : 978-0226843087.

Dekker, S. W. (2003). *When human error becomes a crime*. Human Factors and Aerospace Safety, 3(1):83–92. www.leonardo.lth.se/fileadmin/lusa/Sidney_Dekker/articles/2003_and_before/ErrorCrimeDekker.pdf.

Drupsteen, L. et Guldenmund, F. W. (2014). *What is learning? A review of the safety literature to define learning from incidents, accidents and disasters*. Journal of Contingencies and Crisis Management, 22(2):81–96. **DOI** : 10.1111/1468-5973.12039.

Edmondson, A. C. (1999). *Psychological safety and learning behavior in work teams*. Administrative Science Quarterly, 44(2):350–383. **DOI** : 10.2307/2666999.

EDPS (2025). *Techdispatch 2/2025 — Human oversight of automated decision-making*. Rapport technique, Contrôleur européen de la protection des données. www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-09-23-techdispatch-22025-human-oversight-automated-making_en.

Hazan, É. et Sibony, O. (2026). *Faut-il encore décider ? La décision humaine à l'ère de l'intelligence artificielle*. Flammarion. **ISBN** : 978-2080146687.

Hoff, K. A. et Bashir, M. (2015). *Trust in automation: Integrating empirical evidence on factors that influence trust*. Human Factors, 57(3):407–434. **DOI** : 10.1177/0018720814547570.

Jin, Z. (2025). *Causality for natural language processing*. Preprint arχiv. **DOI** : 10.48550/arXiv.2504.14530.

Leveson, N. (2004). *A new accident model for engineering safer systems*. Safety Science, 42:237–270. sunnyday.mit.edu/accidents/safetyscience-single.pdf.

McMahan, H. B., Moore, E., Ramage, D. *et al.* (2017). *Communication-efficient learning of deep networks from decentralized data*. Dans *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux. **ISBN** : 978-0374257835.

Morrison, E. W. et Milliken, F. J. (2000). *Organizational silence: A barrier to change and development in a pluralistic world*. Academy of Management Review, 25(4):706–725. **DOI** : 10.5465/AMR.2000.3707697.

Nasr, M., Carlini, N., Hayase, J. *et al.* (2023). *Scalable extraction of training data from (production) language models*. Preprint arχiv. **DOI** : 10.48550/arXiv.2311.17035.

Parasuraman, R. et Manzey, D. H. (2010). *Complacency and bias in human use of automation: An attentional integration*. Human Factors, 52(3):381–410. **DOI** : 10.1177/0018720810376055.

Parasuraman, R., Sheridan, T. B. et Wickens, C. D. (2000). *A model for types and levels of human interaction with automation*. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 30(3):286–297.

Pearl, J. (2009). *Causality: models, reasoning and inference*. Cambridge University Press, 2 édition. ISBN : 978-0521895606.

Pearl, J. et Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books. ISBN : 978-0465097609, 432 pages.

Pruckovskaja, V., Weissenfeld, A., Heistracher, C. *et al.* (2023). *Federated learning for predictive maintenance and quality inspection in industrial applications*. Preprint arχiv. DOI : 10.48550/arXiv.2304.11101.

Quiot, A. (2026). *Donneur d'ordre, architecte de la culture de sécurité*. Cahier de la sécurité industrielle 2026-05, Institut pour une culture de sécurité industrielle. À paraître. www.icsi-eu.org.

Rasmussen, J. (1997). *Risk management in a dynamic society: a modelling problem*. Safety Science, 27(2):183–213. DOI : 10.1016/S0925-7535(97)00052-0.

Reason, J. (1997). *Managing the risks of organizational accidents*. Ashgate. ISBN : 978-1840141054, 252 pages.

Rocha, R. (2014). *Du silence organisationnel au développement du débat structuré sur le travail : les effets sur la sécurité et sur l'organisation*. Thèse de doctorat en ergonomie, dirigée par F. Daniellou et V. Mollo, Université de Bordeaux. www.theses.fr/2014BORD0197.

Schein, E. H. et Schein, P. A. (2017). *Organizational Culture and Leadership*. Wiley, 5 édition. ISBN : 978-1119212041, 416 pages.

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books. ISBN : 978-0465068746.

Smith, B. C. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment*. MIT Press. ISBN : 978-0262355209.

Strauch, B. (2018). *Ironies of automation: Still unresolved after all these years*. IEEE Transactions on Human-Machine Systems, 48(5):419–433. DOI : 10.1109/THMS.2017.2732506.

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture and deviance at NASA*. University of Chicago Press. ISBN : 978-0226851754.

de Visser, E. J., Peeters, M. M. M., Jung, M. F. *et al.* (2020). *Towards a theory of longitudinal trust calibration in human-robot teams*. International Journal of Social Robotics, 12(2):459–478. DOI : 10.1007/s12369-019-00596-x.

Wang, D. et Eisner, J. (2017). *Fine-grained prediction of syntactic typology: Discovering latent structure with supervised learning*. Transactions of the Association for Computational Linguistics, 5:147–161. DOI : 10.1162/tacl_a_00052.

Weick, K. E. (1995). *Sensemaking in organizations: Foundations for organizational science*. Sage Publications. ISBN : 978-0803971776, 235 pages.

Weick, K. E. et Sutcliffe, K. M. (2015). *Managing the unexpected: Sustained performance in a complex world*. Wiley, third édition. ISBN : 978-1118862414.

Zečević, M., Willig, M., Dhami, D. S. *et al.* (2023). *Causal parrots: Large language models may talk causality but are not causal*. Transactions on Machine Learning Research. DOI : 10.48550/arXiv.2308.13067.

Vous pouvez extraire ces entrées bibliographiques au format BibTeX en cliquant sur l'icône de trombone à gauche.

Reproduction de ce document

Vous pouvez télécharger ce document, ainsi que d’autres dans la collection des *Cahiers de la Sécurité Industrielle*, depuis le site web de la Foncsi.

FONCSI
Fondation pour une culture de sécurité industrielle

Fondation pour une Culture de Sécurité Industrielle

Fondation de recherche reconnue d’utilité publique

`www.FonCSI.org`

6 allée Émile Monso – CS 22760
31077 Toulouse cedex 4
France

Courriel : `contact@FonCSI.org`

ISSN 2100-3874

FONCSI

Fondation pour une culture
de sécurité industrielle

6 allée Émile Monso
ZAC du Palays - CS 22760
31077 Toulouse cedex 4

www.foncsi.org