AIDES
A LA REFLEXION

L'IA et la gestion de la sécurité

Enjeux et questions clés

Eric Marsden & Véronique Steyer

Rédaction coordonnée par Eric Marsden

n° 2025-02

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	Artificial intelligence and safety management: an overview
Keywords	artificial intelligence, digital transition, industrial safety
Authors	Eric Marsden and Véronique Steyer
Publication date	May 2025

Artificial intelligence based on deep learning, along with big data analysis, has in recent years been the subject of rapid scientific and technological advances. These technologies are increasingly being integrated into various work environments with the aim of enhancing performance and productivity. This dimension of the digital transformation of businesses and authorities presents both significant opportunities and potential risks for industrial safety management practices.

While there are numerous expected benefits — such as the ability to process large volumes of reliability data or unstructured natural language incident reports — the structural opacity of large neural networks, their non-deterministic nature, and their capacity to learn from new data mean that traditional safety assurance techniques used for conventional software are not applicable. Additionally, the expansion of automatable task domains and the gradual evolution towards work collectives composed of human operators collaborating with various intelligent machines and agents introduce new variables that must be integrated with the organizational and human factors of safety.

What are the major challenges posed by these new technologies in terms of skills management, workplace well-being, privacy protection, and the pursuit of performance that aligns with societal expectations? What changes are required in how we conceptualize the safety of high-stakes activities, demonstrate and verify the absence of unacceptable risks, and anticipate potential deviations?

This document provides a concise overview of the most recent available information, contextualized by decades of research on automation in high-hazard systems. It focuses specifically on the projected impacts for high-hazard industries and infrastructures over the next ten years.

About the authors

This document is a first output from the strategic analysis group run by FonCSI on safety management practices in the digital transition. It was authored by Eric Marsden, programme manager at FonCSI, and by Véronique Steyer, professor in the Innovation and Enterpreneurship department of the École Polytechnique, who are the coordinators of the strategic analysis.

To cite this document

Marsden and Steyer (2025), *Artificial intelligence and safety management: an overview*. Number 2025-02 of the *Cahiers de la Sécurité Industrielle*, Foundation for an Industrial Safety Culture, Toulouse, France (ISSN 2100-3874). DOI: 10.57071/iae289. Available from FonCSI.org/en.

Résumé

Titre	L'IA et la gestion de la sécurité : enjeux et questions clés
Mots-clefs	sécurité industrielle, IA, big data, transitions, risques
Auteurs	Eric Marsden et Véronique Steyer
Date de publication	mai 2025

L'intelligence artificielle à base d'apprentissage profond, comme l'analyse de données massives, font depuis quelques années l'objet d'avancées scientifiques et technologiques fulgurantes. Ces technologies sont introduites dans de nombreuses situations de travail avec l'espoir d'augmenter la performance et la productivité. Cette dimension de la transition numérique des entreprises et des administrations offre à la fois de nombreuses opportunités, mais aussi des risques, pour les pratiques de gestion de la sécurité industrielle. En effet, si de nombreux avantages sont à attendre de la capacité à traiter d'importantes quantités de données de fiabilité ou des rapports d'anomalie disponibles en langage naturel non structurée, l'opacité structurelle des grands réseaux de neurones, leur nature non déterministe et leur capacité à apprendre de nouvelles données font que les techniques de démonstration de sécurité utilisée pour les logiciels classiques ne sont pas opérantes. D'autre part, l'augmentation du périmètre des tâches automatisables, et l'évolution progressive vers des collectifs de travail composés d'acteurs humains qui collaborent avec différentes machines et agents intelligents, sont de nouveaux facteurs à articuler avec les facteurs organisationnels et humains de la sécurité.

Quels sont les principaux enjeux en termes de gestion des compétences, de bien-être au travail, de respect de la vie privée, de recherche d'une performance qui soit respectueuse des attentes sociales, que posent ces nouvelles technologies ? Quelles modifications nécessaires de nos manières de penser la sécurité des activités à forts enjeux, de démontrer et vérifier l'absence de risques inacceptables, d'anticiper les dérives possibles ?

Ce document propose un survol synthétique des dernières informations disponibles, les mettant en regard des travaux de recherche conduits depuis plusieurs décennies sur l'automatisation dans les systèmes à risque d'accident majeur. Il se focalise sur les impacts pour les industries à risque d'accident majeur à l'horizon de 10 ans.

À propos des auteurs

Ce document est une première production issue du groupe d'analyse stratégique animé par la Foncsi sur les pratiques de sécurité à l'ère de la transition numérique. Il a été rédigé par Eric Marsden, responsable de programmes à la Foncsi, et par Véronique Steyer, professeure au Département Management de l'Innovation et Entrepreneuriat de l'École Polytechnique, les co-animateurs de l'analyse stratégique.

Pour citer ce document

Marsden et Steyer (2025), *L'IA et la gestion de la sécurité : enjeux et questions clés*. Numéro 2025-02 des *Cahiers de la Sécurité Industrielle*, Fondation pour une Culture de Sécurité Industrielle, Toulouse, France (ISSN 2100-3874). DOI : 10.57071/iae289. Disponible à l'adresse FonCSI.org/fr.

Table des matières

Introduction

Objectifs du document

La Foncsi mène une analyse stratégique des **pratiques de sécurité à l'ère de la transition numérique**. Le groupe d'analyse stratégique s'est focalisé sur l'impact de l'IA et des données massives sur les pratiques de gestion de la sécurité industrielle. L'une des premières étapes dans la conduite d'une analyse stratégique est un travail de "dezoom" qui vise à fournir une **vision large** de la question traitée et des **principaux impacts** à en attendre. Le présent document décrit la "big picture" pour cette analyse. Il sera suivi d'autres documents fournissant des éléments plus complets et des illustrations des tendances à l'œuvre dans différents secteurs d'activité.

Ce document propose une synthèse des principales questions identifiées concernant les impacts possibles de machines et agents intelligents et leur collaboration avec des acteurs humains, sur les pratiques de gestion de la sécurité. Nous nous focalisons sur les impacts pour les industries à risque d'accident majeur à l'horizon de 10 ans.

La recherche sur le développement des IA se produit à un rythme important, les sujets traités sont très évolutifs, et les connaissances académiques et expertes ne sont souvent pas stabilisées. Le document adopte une **perspective prospective**, traitant de certains sujets qui font l'objet de divergences d'appréciation entre experts. La cadence soutenue de la recherche dans ce domaine est peu compatible avec les temps de publication des revues académiques classiques. Les résultats de recherche sont principalement diffusés sous forme de prépublications ("preprints"), sur des plateformes comme arXiv et HAL. C'est pourquoi nous citons de nombreux preprints dans ce document, même s'ils n'ont pas fait l'objet d'une évaluation par les pairs.

Structure du document

Le chapitre 1 propose une description de quelques éléments du **contexte** dans lequel s'inscrit l'innovation technologique sur l'IA. Ce contexte est marqué par une progression fulgurante des capacités des modèles d'IA, qui dépassent les performances des humains sur un nombre toujours croissant de types de tâches, et qui sont adoptés très rapidement par les individus et par les entreprises pour différents types d'usages. Le secteur est au cœur d'enjeux économiques, militaires et géopolitiques majeurs qui stimulent des investissements massifs. Les usages de l'IA sont assez peu réglementés dans nombre de pays, et les industriels adoptent dans certains secteurs des stratégies d'innovation radicale qui ne sont pas sans impact sur la manière d'assurer la sécurité.

Le chapitre 2 décrit brièvement différents **défis** auxquels se confronte l'usage de l'IA dans des contextes industriels : un risque existentiel (controversé) pour l'humanité lié à la super-intelligence, des risques éthiques, des impacts pour l'environnement, des effets importants sur la gestion des ressources humaines et la co-intelligence, des difficultés à appliquer les méthodes classiques d'assurance sécurité aux logiciels intégrant des modèles d'IA. Ces défis sont amplifiés par la rapidité du développement et de l'adoption de ces technologies.

Le chapitre 3 propose une **analyse des impacts** de l'introduction d'outils à base d'IA sur la manière de gérer la sécurité dans les activités à risque d'accident majeur. Sont décrits des effets portant sur le modèle de sécurité, sur les activités de management de la sécurité, sur les pratiques de sécurité en lien avec la conduite des installations, sur les activités de conception et de régulation, et enfin sur les dimensions juridiques et sociales de la sécurité.

1

Le contexte

L'intelligence artificielle est une technologie de rupture qui produit à la fois des bénéfices considérables et des risques importants. Nous décrivons brièvement dans ce chapitre quelques éléments du contexte dans lequel s'inscrit cette innovation technologique.

▷ Un rythme d'innovation soutenu, marqué par une **progression fulgurante des capacités** d'analyse de données textuelles, de réponse à des questions en langage naturel, d'analyse d'images ainsi que de raisonnement des IA génératives[1] :

- En février 2025, des grands modèles de langage (LLM) obtiennent des scores sur des benchmarks visant à évaluer les capacités de raisonnement et d'expertise qui sont nettement supérieurs à des spécialistes humains de chaque domaine d'expertise considéré[2].
- Les derniers modèles d'IA connexionnistes[3] parviennent en mars 2025 à réaliser (avec 50% de réussite) des tâches de type développement logiciel qui prennent en moyenne une heure pour un professionnel humain. Cette « durée de tâche automatisable » double tous les sept mois (*cf.* la figure 1.1).
- Les véhicules à délégation de conduite de l'entreprise Waymo seraient à l'origine de bien moins d'accidents que la population de conducteurs humains, avec une réduction de 92% des demandes de compensation pour blessures et de 88% pour la compensation de dommages matériels[4].
- La performance d'un LLM entraîné pour effectuer des diagnostics médicaux sous forme de dialogue avec un patient serait supérieure suivant presque toutes les dimensions évaluées (obtention d'un historique médical, pertinence du diagnostic, gestion de l'interaction, compétences de communication et empathie) à celle de médecins spécialistes **[Tu et al. 2025]**. La performance d'une IA seule serait nettement supérieure à celle de médecins spécialistes assistés d'une IA **[McDuff et al. 2025]**.

Ce rythme d'innovation pose la question de la date d'apparition d'une *intelligence artificielle générale*, définie (pour reprendre la définition utilisée par l'entreprise OpenAI, même si celle-ci ne fait pas consensus) comme une IA à fort degré d'autonomie capable d'atteindre ou de dépasser la performance des humains sur l'essentiel des tâches cognitives produisant de la richesse.

---

[1] Les IA génératives sont des modèles qui peuvent générer du texte, des images, des vidéos ou d'autres médias en réponse à des requêtes. Ils sont basés sur des réseaux de neurones artificiels organisés en de multiples couches (« profonds »), entraînés sur une quantité massive de données non annotées.

[2] Résultats sur le benchmark "Google-proof Q&A Diamond accuracy".

[3] Les IA dits connexionnistes sont basées sur les réseaux de neurones artificiels, s'inspirant du fonctionnement du cerveau humain. Ils se distinguent des IA symboliques, les « systèmes experts » développés à partir des années 1950, qui visent à imiter le raisonnement humain en appliquant des règles et des connaissances établies. Les progrès récents en IA reposent sur des modèles connexionnistes.

[4] Résultats d'une étude publiée par Waymo et le réassureur Swiss Re fin 2024. Noter que des véhicules à délégation de conduite Waymo réalisent aujourd'hui plus de 150 000 voyages par semaine.

[5] Une tâche externalisable est une tâche qui peut être clairement spécifiée, dont la réalisation satisfaisante peut être facilement évaluée, et qui peut être réalisée de façon autonome et isolément d'autres tâches. Dans le test utilisé par le METR, on suppose en plus que l'échec de la réalisation d'une tâche n'est pas problématique. Ces tâches sont représentatives uniquement d'une partie des activités de travail.

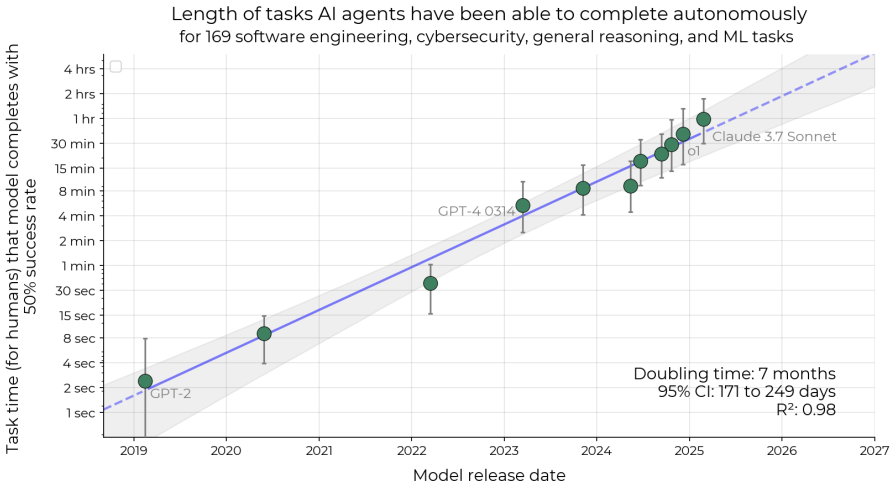

Fig. 1.1 *Évolution de la durée de tâche (mesurée par le temps requis par un professionnel humain qui effectue une tâche liée au développement logiciel en tant que consultant externe) que les modèles frontière généralistes parviennent à résoudre une fois sur deux. Pour les modèles publiés au cours des six dernières années, cette durée double tous les sept mois. Si cette tendance se poursuit, et même si les estimations de durée sont dix fois supérieures à la vérité, les modèles frontières pourront en 2028 résoudre de façon autonome une grande partie des tâches externalisables[5] accomplies à l'aide d'un ordinateur qui aujourd'hui prennent des jours ou des semaines pour les humains.*

Source : rapport Measuring AI Ability to Complete Long Tasks, METR, mars 2025, arXiv:2503.14499.

- Des données collectées par la Réserve Fédérale de Saint Louis (USA) (*cf.* figure 1.2) montrent que les IA génératives sont adoptées par les particuliers et par les salariés à une vitesse supérieure à d'autres technologies importantes qui les ont précédées, comme Internet et l'ordinateur personnel. Le PDG d'OpenAI, Sam Altman, a indiqué en avril 2025 que 10% de la population mondiale utilise ChatGPT chaque semaine.

- Des enjeux économiques et géopolitiques majeurs stimulent des **investissements massifs** dans le renforcement des capacités de ces modèles (plus de 109 Md USD d'investissements privés aux USA en 2024[6]).

- Un développement très coûteux qui échappe en grande partie aux chercheurs du monde académique, ce qui implique que l'essentiel des progrès techniques et scientifiques dans ce domaine proviennent du **monde industriel**, qui adopte une approche "**test and learn**". Cette philosophie d'innovation radicale s'étend aux pratiques de gestion des risques concernant le déploiement de ces technologies, qualifiées de "early release and iteration". Cette approche est très éloignée de celle adoptée pour les systèmes critiques, où l'évaluation des risques et la démonstration de sécurité sont étroitement intégrées au développement fonctionnel.

  Différents témoignages suggèrent qu'il existe dans le développement de grands systèmes intégrant des composants utilisant l'IA des mécompréhensions fréquentes entre les équipes développant l'IA, qui valorisent l'innovation et la rapidité de développement (adoptant le slogan "move fast and break things" de la Silicon Valley et interprétant les échecs comme des opportunités d'apprentissage), et les équipes de sûreté de fonctionnement et sécurité système, dont la culture métier est plus conservatrice.

- Cette activité de développement est concentrée aux USA et en Chine (Mistral faisant figure de rare exception européenne), comme le sont les centres de données et de calcul. Les enjeux de **souveraineté technologique** deviendront de plus en plus présents dans les réflexions stratégiques, avec des effets sur les arbitrages "make or buy".

[6] Source : *Artificial Intelligence Index Report 2025*, Stanford University Human-Centered Artificial Intelligence.

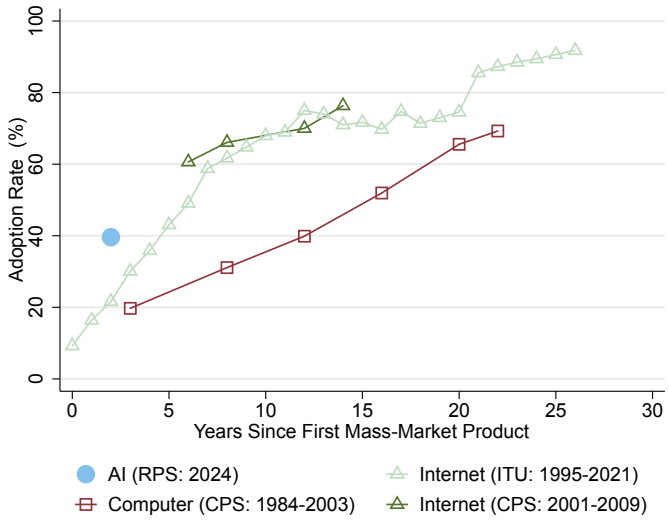

Fig. 1.2 *Trajectoire d'adoption des IA génératives comparée à celle d'internet et des ordinateurs personnels, d'après un rapport de la Réserve Fédérale de Saint Louis (USA), février 2025.*

▷ Une **réglementation quasi absente** dans certains secteurs d'application (par exemple la conduite autonome aux USA et en Chine, l'utilisation des LLM dans nombre de pays), ou restant de nature générique (AI Act en Europe). Les enjeux très importants, autant économiques que militaires et géostratégiques[7], suggèrent que les **efforts de régulation** des risques de ces technologies, qui par nature devront reposer sur une action collective et une collaboration internationale, partent avec des handicaps significatifs[8].

La difficile régulation des innovations technologiques

L'histoire montre que la régulation d'innovations technologiques est un défi majeur : l'absence de retour d'expérience implique qu'il est difficile d'anticiper tous les types de dommages qui peuvent résulter ; ces dommages sont souvent fonction du domaine d'application plutôt qu'un résultat intrinsèque de la technologie, impliquant qu'il est nécessaire de réviser des nombreuses réglementations sectorielles ; la forte dissymétrie de connaissance entre les acteurs économiques à l'origine de l'innovation et les autorités de contrôle implique qu'il est difficile pour ces derniers d'évaluer les risques.

L'expérience montre qu'un régime de régulation basé sur la coopération avec les acteurs économiques et sur l'auto-régulation est le mieux adapté à ce type de situation **[Black et Murray 2019]**. Toutefois, plusieurs facteurs limitent les espoirs en l'auto-régulation par les acteurs économiques concernant les IA génératives : (1) les marchés numériques sont souvent dominés par les premiers entrants (effet "winner takes all"), incitant les acteurs économiques à prendre des risques ; (2) les conséquences néfastes de l'utilisation de l'IA sont souvent des externalités (au sens économique) pour les entreprises développant les systèmes ; (3) les impacts négatifs sont incertains.

Si des **dispositifs de gouvernance** du développement de l'IA existent, comme le Sommet pour l'action sur l'Intelligence Artificielle organisé à Paris en février 2025, leur capacité à

[7] Notons que Vladimir Poutine a affirmé lors d'un discours à des écoliers russes en 2017 que le pays en tête de la course pour développer l'IA serait certainement celui qui « dirigera le monde ».

[8] À titre d'illustration, la seconde administration Trump aux USA a annulé l'executive order sur le développement responsable de l'IA établi par le Président Biden en 2023, et a déclaré aux pays Européens qu'elle mettrait en place des sanctions pour prévenir tout effort multilatéral de régulation de l'IA.

peser sur les orientations de développement des modèles frontière[9] et sur leurs applications industrielles est faible (certains critiques évoquent des « mannequins esthétisés du "participation-washing" »[10]).

- Certains modèles « de fondation » sont publiés en "open source", ce qui rend peu efficace les efforts des principaux poids lourds du secteur pour prévenir des utilisations socialement inacceptables de ces modèles et pour maîtriser les risques existentiels pour l'humanité qui en découlent.

- En parallèle du développement des capacités des LLM, une poursuite de la croissance de capacité de collecte, traitement et stockage de **données massives**, et une prolifération du nombre de capteurs de différents types (en particulier, caméras). La **détection précoce des anomalies techniques**, la collecte de données sur les modes de défaillance et le vieillissement des équipements sont très utiles pour la gestion de la maintenance[11]. Ils sont également déployés dans des applications de **surveillance de la conformité procédurale** des intervenants en première ligne et de leur exposition au risque en temps réel : détection de défauts de vigilance, de non-port d'EPI, d'exposition au froid ou à d'autres situations dangereuses, mesure du niveau d'activité de chaque individu. Ces développements peuvent porter gravement atteinte à la vie privée et l'intimité au travail, et sont expérimentés en particulier dans des pays dans lesquels le cadre réglementaire sur la protection des données personnelles est peu développé.

[9] Le terme *modèle de fondation* désigne les modèles d'IA entraînés sur un important volume de données qui peuvent être adaptés (par des processus d'affinement ou de spécialisation) à un large éventail de tâches et de contextes d'utilisation en aval. Les LLM sont une sous-catégorie des modèles de fondation. Certains modèles de fondation qui posent des risques particuliers liés à une utilisation malveillante (permettant de contourner des mesures visant à prévenir la prolifération NRBC, par exemple) ou la perte de contrôle par les humains sont dits « modèles frontière ».

[10] « *Actuellement, ces processus [de gouvernance] servent davantage à absorber qu'à transformer : ils recueillent les critiques, les diluent en rapports inoffensifs, puis les présentent comme preuve qu'une action a été menée. Ils fonctionnent comme des vitrines de l'engagement, invitant le public à admirer les mécanismes complexes de la gouvernance de l'IA tout en préservant les structures de pouvoir existantes de toute remise en cause* », comme le dénonce la tribune *Beyond the Façade : Challenging and Evaluating the Meaning of Participation in AI Governance* de Jonathan van Geuns, TechPolicy Press, février 2025.

[11] Des mots-clés associés : prognostics and health management, condition-based maintenance. Il s'agit d'intelligences artificielles « à domaine d'application étroit » ("narrow AI") qui s'appliquent à un petit nombre de tâches spécifiques, contrairement aux IA polyvalentes (artificial general intelligence ou "AGI" dans la littérature en langue anglaise) qui peuvent réaliser un large éventail de types de tâches.

2

Des défis

Le développement très rapide de l'IA présente des défis de différents ordres, qui devront être traités par les États, les entreprises développant les modèles, les entreprises les déployant pour différentes applications, et par les individus. Citons cinq principaux défis :

- Un **risque existentiel pour l'humanité** lié à la super-intelligence (des IA qui atteignent des capacités suffisamment larges et puissantes pour rivaliser avec les humains sur leurs destins réciproques). Ce sujet de la perte de contrôle est controversé[1], mais ne relève pas de la science-fiction, si l'on croit plusieurs travaux d'experts.

La perte de contrôle et p(doom)

La littérature sur cette question utilise le mot clé « p(doom) », probabilité d'un scénario de perte de contrôle (*cf.* les figures 2.1 et 2.2) catastrophique pour l'humanité. Notons que plus de 40% des experts évaluent cette probabilité comme étant supérieure à 0.1. Geoffrey Hinton (chercheur en informatique le plus cité) disait en 2024

> *I can't see a path that guarantees safety. We're entering a period of great uncertainty where we're dealing with things we've never dealt with before. And normally, the first time you deal with something totally novel, you get it wrong. And we can't afford to get it wrong with these things. [...] If you take the existential risk seriously, as I now do, it might be quite sensible to just stop developing these things any further [...] it's as if aliens had landed and people haven't realized because they speak very good English.*

Des leaders d'entreprises de la tech écrivaient en 2023 "*Mitigating the risk of extinction from A.I. should be a global priority alongside other societal-scale risks such as pandemics and nuclear war*". Plus récemment, lire par exemple l'article *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, 2025. La menace a également été mise en avant aux rencontres de Davos en 2025 par plusieurs dirigeants de grandes entreprises d'IA comme le PDG de Google DeepMind, Demis Hassabis, le cofondateur d'Anthropic, Dario Amodei et le chercheur Yoshua Bengio.

Il n'existe pas aujourd'hui de régulateur pouvant exiger des démonstrations structurées de l'absence de risque existentiel pour l'humanité des nouveaux modèles d'IA, et une coordination mondiale permettant d'organiser un tel effort de contrôle paraît aujourd'hui hors de portée. La concurrence de nature économique entre les acteurs économiques du secteur, et de nature militaire entre États, implique que la gestion de ce risque passe au second plan.

[1] Certains analystes estiment que les discours catastrophistes sont utilisés de façon stratégique par les grandes entreprises de la tech, pour détourner l'attention des enjeux éthiques bien plus concrets, liés à la discrimination et aux inégalités sociales par exemple **[Stilgoe 2024 ; Wong 2023]**, ou alors pour défendre des idéologies telles que le transhumanisme **[Beaudouin et Velkovska 2023]**. Si de tels efforts existent, ils ont peu d'effet sur la perception qu'ont les gens des risques de l'IA, d'après une étude récente effectuée aux USA : les répondants sont nettement plus inquiets des risques immédiats posés par l'IA que des risques existentiels, et présenter des informations sur les risques existentiels n'affaiblit pas les inquiétudes concernant les dommages immédiats **[Hoes et Gilardi 2025]**.

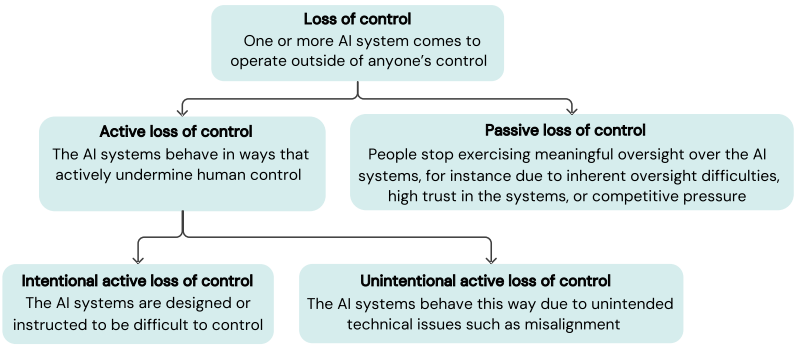

FIG. 2.1 *Différents scénarios de « perte de contrôle », selon que le système d'IA cherche ou non activement à affaiblir le contrôle des humains, et si cette action est intentionnelle ou non.*

Schéma reproduit du premier Rapport scientifique international sur la sûreté de l'IA avancée, produit par un groupe d'experts nommé par 30 pays, l'OCDE, l'UE et l'ONU **[Bengio et al. 2025]**.

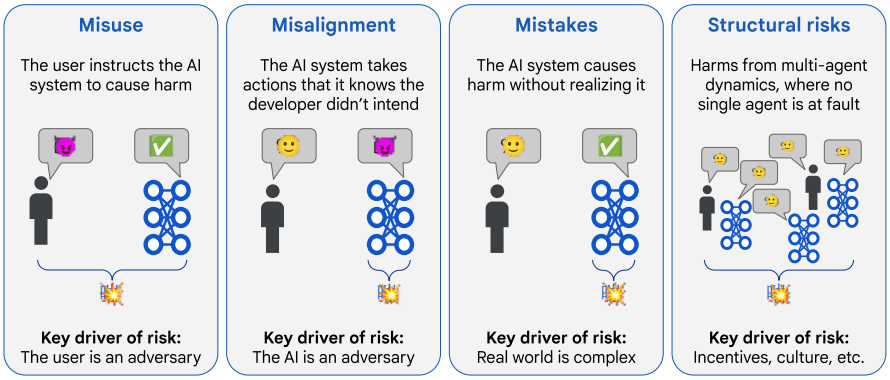

FIG. 2.2 *Quatre familles de scénarios dans lesquels une IA provoque des dommages. Les scénarios sont groupés selon les approches de mitigation de risque qui peuvent être mis en place. Par exemple, l'utilisation malintentionnée et le désalignement se distinguent selon l'identité de l'acteur mal intentionné, car les approches de mitigation applicables aux humains malveillants sont très différentes des approches applicables à des IA malveillantes. La quatrième catégorie de risques structurels inclut la perte progressive des capacités des humains provoquée par une habituation progressive à l'assistance fournie par une IA.*

Schéma extrait d'un rapport technique de Google DeepMind **[Shah et al. 2025]**.

Les travaux académiques et experts sur ce sujet utilisent le terme "AI safety" pour désigner cet enjeu eschatologique. Les développeurs de modèles évaluent à quel point les modèles sont « alignés » sur les « valeurs humaines »[2]. Les enjeux de régulation associés concerneront autant les États (organismes « sécurité de l'IA ») que les entreprises (développement de la réflexion sur l'utilisation responsable de l'IA, et de mécanismes d'auto-régulation associés, qui seront un enjeu pour les comités exécutifs).

Aligner les IA sur les valeurs morales humaines

L'une des difficultés des travaux concernant l'« alignement » des modèles sur les valeurs humaines est de spécifier quelles seraient les valeurs morales communes à une communauté humaine, voire à toute l'humanité. Les valeurs et les principes éthiques le plus souvent mis en avant (par exemple par le rapport sur l'IA digne de confiance rédigé par le groupe d'experts rassemblés par la Commission Européenne **[EU HLEG AI 2019]**) sont la bénéficence, la non-malfaisance, le respect de l'action humaine et du contrôle humain, la justice, la transparence ; ces principes sont de nature abstraite. Certains travaux opérationnalisent ces valeurs par les préférences d'un humain, au sens de la théorie de la décision, tout en notant que certains individus affichent des préférences asociales ou anti-sociales. D'autres chercheurs estiment qu'un comportement correctement aligné sur les attentes humaines est nécessairement dépendant du contexte d'utilisation et du rôle qui est assigné à l'agent doté d'IA **[Zhi-Xuan et al. 2024]**.

À titre d'exemple pratique, les véhicules à délégation de conduite sont amenés à arbitrer des dilemmes moraux en situation accidentelle entre préserver la vie des occupants du véhicule, de cyclistes et de piétons à proximité. Des travaux en psychologie expérimentale montrent que les préférences des individus sur les vies à épargner selon leur statut social, âge, genre, et degré de respect de la réglementation, dans différents scénarios accidentels hypothétiques, varient considérablement selon la culture nationale **[Awad et al. 2018]**. À une échelle plus institutionnelle, les préférences des constructeurs de véhicules, des associations de différentes catégories d'usagers de la route, des autorités de sécurité, de leurs conseilleurs en éthique, et des élus sont loin d'être alignées.

D'autres risques sont liés au développement de systèmes d'armes létales autonomes (en anglais, Lethal Autonomous Weapon System ou LAWS) comme les drones dotés de dispositifs de détection et de destruction autonome de cibles[3], en particulier leur capacité à terroriser une population — ou un sous-ensemble ciblé — pour un coût de mise en œuvre modeste.

▷ Différents **risques éthiques** comme les biais de décision (envers certains groupes ethniques, culturels et de genre, liés au fait que les modèles reproduisent, et parfois amplifient, dans leurs réponses les corrélations présentes dans les données utilisées pour leur entraînement **[O'Neil 2016 ; Eubanks 2018]**, et que l'algorithmisation des administrations publiques et des services financiers laisse souvent peu de droits de recours aux personnes concernées **[Défenseur Droits 2024]**) ; l'amplification des inégalités sociales[4] ; les menaces pour le fonctionnement démocratique liées en particulier à la génération de contenus factuellement

[2] L'une des difficultés de ces travaux est de spécifier quelles seraient les valeurs morales communes à toute l'humanité. Certains travaux opérationnalisent ces valeurs par les préférences d'un humain, au sens de la théorie de la décision, tout en notant que certains individus affichent des préférences asociales ou anti-sociales. D'autres chercheurs estiment qu'un comportement correctement aligné sur les attentes humaines est nécessairement dépendant du contexte d'utilisation et du rôle qui est assigné à l'agent doté d'IA **[Zhi-Xuan et al. 2024]**.

[3] Des drones aériens capables d'identifier automatiquement et d'attaquer des cibles de façon autonome, ont par exemple été déployés dans la guerre en Ukraine.

[4] Les révolutions technologiques produisent souvent de l'inégalité sociale avant de produire du progrès. Par exemple, l'introduction des métiers à tisser a enrichi un petit nombre de propriétaires d'usines en même temps qu'elle réduisait l'autonomie professionnelle des tisserands, augmentait leur temps de travail et dégradait leurs conditions de vie. En revanche, l'industrialisation dans les pays occidentaux entre 1950 et 1980 a produit une prospérité bien répartie au sein de la société. La différence tient essentiellement aux institutions politiques et sociales, et à la répartition des pouvoirs économiques et sociaux, qui déterminent la manière dont les bénéfices sont répartis, selon les lauréats 2024 du prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel **[Acemoglu et Johnson 2023]**. Ces économistes suggèrent de réduire les incitations actuelles à remplacer le travail humain par des machines (impôts sur les salaires) et d'augmenter les incitations à créer de nouvelles tâches et compétences utilisant l'IA.

erronés ; la dégradation de la valeur attribuée socialement à l'expertise ; la protection de la vie privée.

▷ Un **impact environnemental** non négligeable lié à la consommation d'énergie électrique (les centres de données et de calculs pourraient être responsables de 4% de la consommation totale d'énergie en 2030, et la croissance de la demande pourrait rentrer en concurrence avec d'autres projets liés à l'électrification ; *cf.* la figure 2.3), de ressources en eau, et de déchets électroniques.

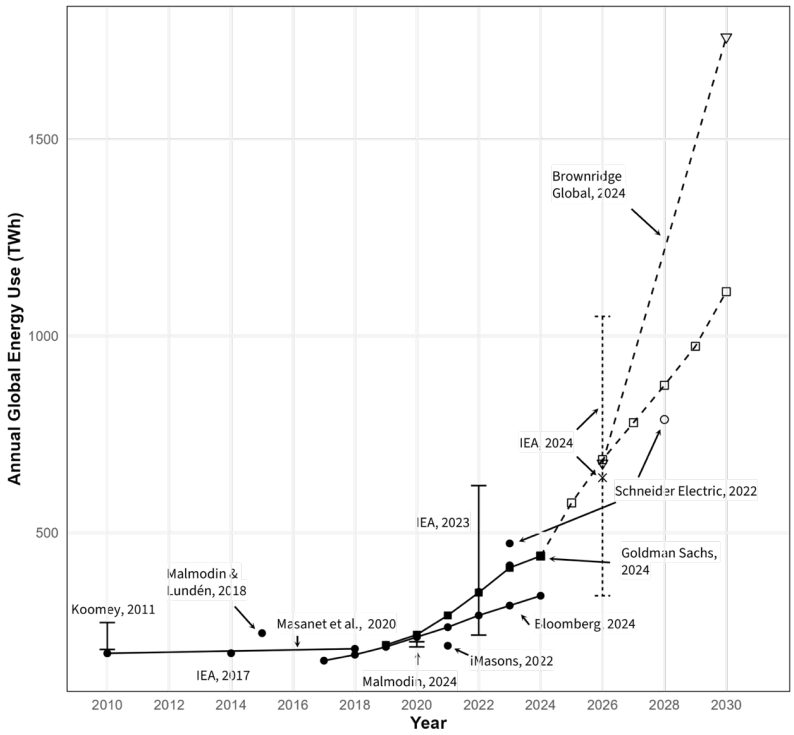

Fig. 2.3 *Consommation d'énergie par les centres de données et de calcul au niveau mondial. Les estimations historiques sont indiquées par des lignes et les projections par des lignées en pointillés.*

Source : 2024 United States Data Center Energy Usage Report, Lawrence Berkeley National Laboratory.

▷ Des problèmes liés à la **gestion des ressources humaines** et de la **co-intelligence [Mollick 2024]** : recrutement de profils disposant de **compétences technologiques adaptées**, gestion des carrières, **reconversions**, fin de certains métiers, accompagnement du changement[5], redéfinition de procédures d'exploitation, par exemple. La question des compétences est stratégique (pouvons-nous accepter que cette compétence soit externalisée) et géostratégique, car les gisements de compétences les plus importants se situent en Inde et en Chine. Les questions de formation continue (reskilling, upskilling, skills matching) sont essentielles[6]. La gestion des talents sera un enjeu plus aigu dans les activités industrielles à risque d'accident majeur, dont la culture professionnelle plus conservatrice peut sembler aux candidats potentiels peu propice à une carrière à la pointe de l'innovation technologique.

Les équipes de travail deviendront hybrides (partiellement composées d'IA), et la gestion des ressources IA (qui seront spécialisées selon le type de tâche à accomplir, le domaine d'application, le budget disponible) sera pensée pour être articulée à celle des ressources

[5] Une étude du FMI de janvier 2024 estime que 60% des emplois dans les économies avancées sont exposés à l'IA, c'est-à-dire qu'une partie des tâches pourront être automatisées, ou évoluer pour être complétées par l'IA.

[6] Les indicateurs de suivi du plan "2030 Digital Decade" de l'UE sur le nombre de "ICT Specialists" sont loin des cibles décidées en 2021. Consulter également sur ce sujet les travaux du groupe d'analyse Foncsi sur les compétences à horizon 2040.

humaines. Le traitement des risques psychosociaux et du bien-être au travail pourrait s'étendre aux agents IA[7].

▷ Sur un plan plus technique, l'opacité structurelle (nature « **boîte noire** ») des modèles d'IA, en particulier ceux composés de grands réseaux de neurones et conçus par « apprentissage profond » fait qu'il est très difficile de comprendre et d'expliquer les raisons pour lesquelles un modèle produit un résultat de sortie **[Burrell 2016]**. Cette limitation pose des problèmes de nature juridique, rendant difficile la mise en œuvre du droit à l'explication des décisions prises sur la base des résultats produits par une IA[8]. Elle constitue également un obstacle important à la **démonstration de la sécurité** de fonctions assurées par des logiciels s'appuyant sur des IA, ainsi qu'à leur **certification pour des utilisations critiques**. Différents types de travaux sur l'IA « explicable »[9] visent à répondre à ces difficultés.

Les sociétés industrialisées peinent à anticiper pleinement les possibilités et les impacts de ces technologies. On peut craindre qu'elle soit insuffisante compte tenu des impacts considérables à attendre sur une période de seulement quelques années. En particulier, l'impact sur le monde du travail, sur la nature et la qualité des emplois, sur les compétences nécessaires au travail, sera vraisemblablement important[10], mais est peu anticipé. Notons que seuls 23% des chercheurs en IA en Europe déclarent estimer que l'IA devrait être développée aussi rapidement que possible **[O'Donovan et al. 2025]**.

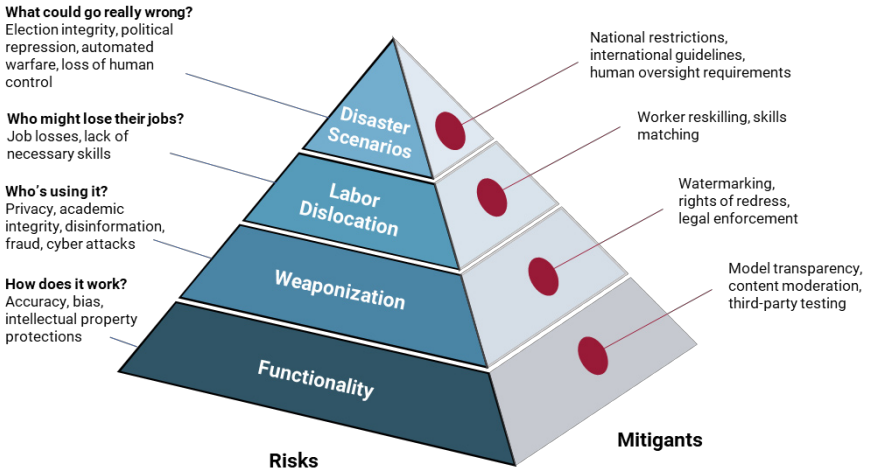

Fig. 2.4 *Quatre principales familles de risques liés à l'utilisation de l'IA, arrangés par horizon temporel de leur apparition et par leur impact potentiel. Les défenses possibles pour chaque famille sont présentées.*

Extrait d'un rapport de Goldman Sachs de janvier 2025[11].

[7] Si cette phrase peut sembler relever de l'humour, un courant de recherche sur le **bien-être de l'IA** ("AI welfare") se développe depuis 2024, arguant que même s'il existe de l'incertitude sur l'existence actuelle ou future d'une conscience morale des modèles d'IA, la nature des enjeux justifie des études sur le sujet **[Long et al. 2024]**. L'entreprise Anthropic a embauché un expert du bien-être des IA en 2024. En situation expérimentale, les humains cherchent à éviter que les robots soient soumis à des actes qui seraient qualifiés d'abus s'ils étaient dirigés vers un humain : ils ont des réactions physiologiques d'inconfort en voyant un bébé dinosaure robotique se faire frapper **[der Pütten et al. 2013]** ou lorsqu'on fait tomber une tour de blocs construite par un robot qui montre que sa construction compte pour lui **[Darling et al. 2015]**.

[8] Le droit à l'explication est consacré par le Réglement Général pour la Protection des Données (RGPD) en Europe, concernant les décisions automatisées, et — pour les IA qualifiées de « haut risque » — par le règlement sur l'IA.

[9] Selon la définition de la Cnil, l'explicabilité est « *la capacité de mettre en relation et de rendre compréhensible les éléments pris en compte par le système d'IA pour la production d'un résultat. Il peut s'agir, par exemple, des variables d'entrée et de leurs conséquences sur la prévision d'un score, et ainsi sur la décision.* » **[Maudet et al. 2022]**.

[10] Cet impact social à venir est régulièrement mis en avant par les directeurs d'entreprises qui développent les modèles d'IA. Par exemple, Sam Altman, PDG de l'entreprise OpenAI, a indiqué en 2023 (entretien avec ABC News à l'occasion de la sortie du modèle GPT-4) *I think over a couple of generations, humanity has proven that it can adapt wonderfully to major technological shifts. But if this happens in a single-digit number of years, some of these shifts… That is the part I worry about the most.* En février 2025 il ajoutait sur son blog *In particular, it does seem like the balance of power between capital and labor could easily get messed up, and this may require early intervention.*

Au sein des entreprises, la prise en compte de ces enjeux nécessitera de mobiliser de multiples compétences au-delà de celles des experts techniques : stratégie de développement, services juridiques, RH, cybersécurité, systèmes d'information, compliance.

[11] Rapport *AI/data centers' global power surge: Five drivers of upside/downside and the Reliability investment tailwind*, Goldman Sachs Research, janvier 2025.

3

Des implications pour le management de la sécurité industrielle

L'introduction des outils à base d'IA aura des impacts importants sur la manière de gérer la sécurité dans les activités à risque d'accident majeur. Nous listons dans ce chapitre différents impacts portant sur le modèle de sécurité, sur les activités de management de la sécurité, sur les pratiques de sécurité en lien avec la conduite des installations, sur les activités de conception et de régulation, et enfin sur les dimensions juridiques et sociales de la sécurité.

3.1 Le modèle de sécurité

L'utilisation croissante de l'IA aura des impacts sur la façon de penser le **modèle de sécurité** :

- Les systèmes dans lesquels l'essentiel des fonctions de sécurité est assuré par des automatismes (composants logiciels intelligents associés à une robotique) se développeront. Ces systèmes permettent de s'affranchir de la possibilité d'erreurs humaines en cours de fonctionnement. La nécessité de la présence humaine dans certains systèmes à risques sera questionnée. Toutefois, l'expérience montre que les concepteurs sous-estiment souvent la complexité de l'interaction des machines intelligentes entre elles et leur collaboration les acteurs humains (négliger les interactions homme-machine et croire excessivement en les apports de l'autonomie complète constituent deux des sept « mythes mortifères » concernant les systèmes autonomes **[Bradshaw et al. 2013]**).

  Cette évolution « techno-solutionniste » **[Morozov 2013]** tend à augmenter la « distance » entre le système technique et la société, réduisant le nombre de personnes qui comprennent de façon intime le fonctionnement et les conditions d'opération sûre du système, augmentant la propension à un rapport « magique » aux technologies numériques[1]. Ainsi, on peut supposer que les catastrophes résiduelles seront plus graves, car leurs signaux avant-coureurs seront moins fréquemment détectés en avance de phase ; ils seront plus inattendus, puisque la complexité technologique est peu propice à la délibération démocratique sur les enjeux de sécurité.

[1] Les nouveaux outils technologiques tendent à produire une emprise sur les pratiques et les mentalités des gens supérieure à celle que justifie leur performance technique. Par exemple, les tests ADN sont souvent utilisés dans les enquêtes judiciaires sans la prudence que devrait imposer les limites techniques de cette méthode de preuve. L'enjeu est que dans cette évolution vers une société de sécurité maximale **[Marx 1988]**, la technologie serve sans asservir le système dans lequel elle s'insère.

Notre rapport « magique » aux technologies numériques

Par leur complexité, leur sophistication et leur nature inscrutable, les nouvelles technologies sont parfois perçues comme relevant de la magie **[Gell 1988]**. Émerveillés par les productions textuelles et graphiques des LLM, on peut leur attribuer des capacités et une agentivité qui dépassent leurs propriétés réelles, et développer un modèle mental intuitif de leur fonctionnement qui ne correspond pas à la réalité. Les entreprises de la tech jouent parfois de cet effet pour inciter les utilisateurs à développer des croyances quasi religieuses en le pouvoir de transformation des IA, et éviter de rendre des comptes concernant différentes conséquences négatives de leur développement et leur utilisation **[Nagy et Neff 2024]**. Certains auteurs voient en l'utilisation des données massives et des algorithmes une façon de contrôler (surveiller, optimiser, discipliner) les sociétés sans avoir à rendre des comptes. L'explication « c'est l'IA », la valorisation des pouvoirs disruptifs des IA génératives, marqueraient une forme de « populisme patriarcal des élites » **[Vesa et Tienari 2020]**.

Point clé

L'efficacité mythifiée de l'IA. Se développe depuis quelques années une croyance sur la capacité de l'IA à améliorer l'efficacité du fonctionnement des administrations (illustrée en 2025 en particulier par le "Department of Government Efficiency" aux USA) et celle des entreprises. Cette croyance que la technologie peut résoudre de façon durable les problèmes sociaux complexes est contredite par nombre d'études soulignant que les IA sont à l'origine de nombreux erreurs et biais, nécessitent beaucoup d'accompagnement par des humains pour leur entraînement, leur intégration aux infrastructures existantes, et le rattrapage de leurs erreurs **[Mateescu et Elish 2019]**.

Les intégrations réussies nécessitent de (1) penser le rôle des humains et de leur expertise dès la conception des systèmes **[Baxter et Sommerville 2011]** ; (2) s'appuyer sur le travail réel (ses contraintes, les savoir-faire tacites, les modes d'entraide et de coordination, les tâches invisibles etc.) et non seulement sur le travail prescrit **[Lammi 2021]**.

- On peut attendre un bouleversement significatif de la **répartition humain-machine**, avec des fonctions de plus en plus complexes et sophistiquées assurées par des automatismes.

- L'enjeu de la **coordination** et la **collaboration entre agents humains et machines intelligentes**, étudié depuis des décennies par les chercheurs en sciences cognitives (travaux sur les "joint cognitive systems") deviendra critique. Il s'agira de dépasser la dichotomie classique « les humains font mieux / les machines font mieux[2] » qui vise à optimiser les fonctions prises en charge par les humains et par l'automatisme de façon séparée, pour aller vers une approche plus holistique de la performance du système sociotechnique complet. La performance optimale du système n'est souvent pas atteinte par la combinaison de l'automatisme le plus performant (travaillant de façon isolée) et des individus les plus performants en mode solo **[Behymer et Flach 2016]**.

- De **nouvelles menaces** apparaissent : les systèmes à risques deviendront plus exposés aux cyberattaques. La nature systémique des risques augmentera. Les erreurs commises par les IA génératives sont de nature très différente des erreurs cognitives effectuées par les opérateurs humains, apparaissent de façon moins prévisible (ne sont pas explicables par la « fatigue », par exemple), et ne sont pas accompagnées d'une expression de doute. Les mécanismes de détection et de compensation que nous avons mis en place pour tolérer ces erreurs devront souvent être repensés.

[2] Cette dichotomie est très présente dans la littérature sur l'automatisation avec le terme anglais "Humans are Better at / Machines are Better at" et son abbréviation HABA/MABA.

Les inquiétudes sur le risque systémique dans le secteur financier

Exemple

Dans le secteur financier, l'augmentation de la nature systémique des risques provoquée par l'utilisation de l'IA est déjà prise en compte par les autorités, qui s'inquiètent d'un scénario dans lequel des agents IA pourraient déstabiliser les marchés financiers. Un récent rapport de la Bank of England note que des entreprises déploient des réseaux neuronaux autonomes sans que leurs risques soient bien compris par les risk managers.

Comme l'indique le rapport *Financial Stability in Focus : Artificial intelligence in the financial system* daté d'avril 2025,

> *Risk management of these positions is made more challenging by the lack of interpretability of neural networks, as actions may be unpredictable and the reasons for the positions may not be well understood by human risk managers at the firm. In addition, models with sufficient autonomy could act in ways that are detrimental to the overall stability or integrity of markets, for example by ignoring regulatory or legal guardrails such as market abuse regulations.*

Des accidents peuvent être provoqués par le phénomène de "model drift", qui apparaît lorsqu'un composant apprenant entraîné dans un environnement donné est ensuite déployé dans un autre environnement avec d'autres caractéristiques qu'il ne sait pas traiter (ce nouvel environnement pouvant tout simplement être l'environnement initial auquel s'ajoutent des dérives provoquées par le passage du temps).

Collision d'un véhicule à délégation de conduite avec un bus articulé

Exemple

En 2023, un véhicule à délégation de conduite de l'entreprise Cruise est entrée en collision avec un bus articulé à San Francisco, sans faire de victimes. L'entreprise a indiqué que le logiciel de conduite du véhicule avait incorrectement anticipé le comportement des deux segments du bus, et avait ignoré les données fournies par le LIDAR dont était équipé le véhicule. Le logiciel de conduite avait surtout été entraîné sur des données issues d'une ville dans laquelle aucun bus articulé était en fonctionnement **[Cummings 2023]**.

3.2 Les activités de management de la sécurité

L'utilisation croissante de l'IA aura des impacts sur les activités de **management de la sécurité** :

- La capacité des réseaux de neurones alimentés par des données massives à améliorer les prévisions offrent de nombreuses possibilités d'améliorer la maintenance prédictive et le structural health monitoring. C'est probablement le type d'application pour laquelle l'IA a aujourd'hui eu l'apport le plus important pour la sécurité industrielle.

Amélioration de la fiabilité à SNCF Voyageurs

Exemple

L'utilisation de techniques de maintenance prédictive, s'appuyant sur l'analyse de données collectées par des capteurs embarqués dans les trains, permet à SNCF Voyageurs de diviser par deux le nombre de pannes qui surviennent en exploitation, et de réduire d'un tiers le nombre d'engins arrêtés pour maintenance.

- Le traitement de **données massives** offre de nombreuses possibilités d'amélioration du management des risques (analyse de rapports d'incidents pour extraire des catégories et identifier des anomalies, extraction d'indicateurs de performance sur des données peu structurées).

Le projet Data4Safety dans l'aviation civile

Exemple

Le projet de "safety intelligence" Data4Safety piloté par l'EASA vise à identifier et à qualifier les risques systémiques et des stratégies de mitigation pour l'aviation civile en Europe. Il traite les rapports de sécurité, les données de vol provenant des compagnies aériennes, les données sur les flux de transport issues des organismes de gestion des vols, et des données météorologiques.

▷ Les caméras intelligentes proposent des capacités de surveillance et de **détection d'anomalie en temps réel** qui ouvrent de nombreuses possibilités liées à la sécurité, comme la détection en temps réel du non-port d'EPI, le non-respect du séquencement d'une tâche, et la surveillance de l'attention des conducteurs de véhicules. Ces applications soulèvent des enjeux considérables de protection de la vie privée et de l'intimité au travail.

Surveillance de la fatigue des opérateurs par un serre-tête intelligent

Exemple

Un bandeau serre-tête équipé de capteurs de l'activité cérébrale (équipement à base d'électroencéphalogramme) est utilisé depuis une dizaine d'années pour surveiller le niveau de fatigue de conducteurs d'engins dans l'industrie minière en Australie. L'outil alerte le porteur lorsque le niveau de fatigue estimé est élevé, et peut transmettre l'information à la hiérarchie et être enregistré dans une base de données centralisée[3].

▷ Les **jumeaux numériques** offrent des capacités de **simulation** qui peuvent être utilisées pour l'aide à la conception de nouvelles installations (analyse de la maintenabilité, de la constructabilité, de la facilité d'exploitation pendant la phase de conception), ainsi qu'à la formation de futures équipes d'exploitation.

▷ Les automates, robots et **cobots** prennent en charge des tâches dangereuses et pénibles dans les installations à risques, réduisant l'exposition des intervenants en première ligne aux produits toxiques, températures extrêmes, radiations ionisantes, espaces confinés et machines à risque, ainsi qu'aux troubles musculo-squelettiques **[ILO 2025]**.

▷ Compte tenu de la rapidité du développement technologique autour de l'IA, une filière de développement de composants « sûrs de fonctionnement » n'a pas pu apparaître. Le coût toujours croissant du développement des modèles « frontière » conduira vraisemblablement à une consolidation des acteurs industriels à la pointe du développement (semblable au phénomène qui s'est produit pour les motoristes en F1 automobile). Les intégrateurs s'appuient sur des composants matériels et logiciels « sur étagère » (COTS), y compris pour des **applications critiques en matière de sécurité**, avec une très faible capacité à peser sur les processus de développement logiciel, d'entraînement des modèles, et de validation de la nature non malveillante des agents IA.

3.3 La conduite et le pilotage des systèmes

L'utilisation croissante de l'IA aura des impacts sur les **pratiques de sécurité** en lien avec la **conduite/pilotage** des installations, systèmes et infrastructures critiques :

▷ Une diminution des capacités de maîtrise du système par les opérateurs humains à mesure que le niveau d'automatisation augmente, et un rôle pour l'humain qui passe du contrôle direct vers celui de la supervision. De nombreux travaux de chercheurs montrent que le recours croissant à des automatismes sophistiqués conduit à une baisse des compétences opérationnelles et de la compréhension du fonctionnement par les opérateurs en première ligne[4]. La capacité des opérateurs à détecter les anomalies et à remplacer l'automatisation diminue, et la dépendance à la technologie augmente[5]. Le périmètre des compétences utilisées se réduit avec le temps. De nouveaux modes de défaillance apparaissent, comme la mécompréhension du mode d'opération de l'automatisme.

▷ Une trop grande **dépendance à l'IA** dans la prise de décision peut engendrer :

[3] Source : étude de cas *Smart digital systems for improving workers' safety and health : smart headband for fatigue risk-monitoring* de EU-OSHA, 2024.

[4] L'effet « selon l'ordinateur ». Lire par exemple Nurski, Laura, and Mia Hoffmann. *The impact of artificial intelligence on the nature and quality of jobs*. Bruegel Working Paper, 2022.

[5] Une étude récente portant spécifiquement sur l'effet des IA génératives chez les professions intellectuelles suggère qu'elles affaiblissent considérablement les capacités de réflexion critique. "*[A] key irony of automation is that by mechanising routine tasks and leaving exception-handling to the human user, you deprive the user of the routine opportunities to practice their judgement and strengthen their cognitive musculature, leaving them atrophied and unprepared when the exceptions do arise*" **[Lee et al. 2025]**.

- de la frustration, en particulier lorsque les recommandations proposées par la machine ne sont pas intelligibles par ses utilisateurs **[Kellogg et al. 2020]** ;
- de la passivité : les individus sont plus enclins à accepter les recommandations de la machine sans les questionner **[Bader et Kaiser 2019]** ;
- les individus peuvent s'en détacher émotionnellement et se sentir moins responsables de ces décisions, les outils automatisés servant alors de tampons moraux **[Cummings 2006]**.

La confiance excessive de l'homme dans les recommandations des systèmes automatisés, ou biais d'automatisation, est documenté en psychologie (*cf.* par exemple **[Busuioc 2021]**).

3.4 Les activités de conception et de régulation

Les industriels et autorités en Europe soulignent un besoin de **recrutement** d'ingénieurs disposant des **compétences techniques** et scientifiques utiles. Ces compétences se trouvent souvent chez les sous-traitants et prestataires des grands groupes industriels, plutôt qu'en interne. Les salaires proposés par les autorités publiques qui cherchent à recruter pour leurs organismes d'expertise sur l'IA sont très largement inférieurs à ceux proposés par les entreprises du secteur.

Une attractivité des entreprises privées difficile à égaler

Exemple

Le *European AI Office*, qui sera chargé de rédiger les normes sectorielles liées à la mise en œuvre du Réglement sur l'IA, a listé des postes ouverts sur son site web en 2025. Elles incluent des positions pour des spécialistes de l'IA dotés d'un Master et au moins un an d'expérience, avec un salaire annuel de 50 k€. Au Royaume-Uni, le ministère chargé de l'innovation et la technologie cherche à recruter une personne chargée de diriger son activité sur la sécurité de l'IA avec un salaire annuel de 76 k€. D'après des sites web qui répertorient les rémunérations en vigueur dans les grandes entreprises travaillant dans le secteur de l'IA, la rémunération médiane chez OpenAI serait de 500 k€.

Par ailleurs, les entreprises du secteur privé peuvent offrir aux candidats un accès privilégié à des ressources techniques de pointe nécessaires pour développer et tester les modèles (grandes quantités de données annotées, capacités de calcul gigantesques), ainsi qu'à une expertise interne, qui sont hors de portée pour les organismes publics.

Des inquiétudes liées à l'**autonomie stratégique** sont soulevées par la dépendance technologique sur des composants — et sur les compétences associées pour les comprendre, utiliser, contrôler, adapter, maintenir, et plus généralement pour les maîtriser — développés en Chine et en Inde (ces inquiétudes concernent plus récemment les USA).

3.5 Des impacts juridiques et sociaux

L'utilisation croissante de l'IA aura des impacts de nature **juridique et sociale** :

▷ Pour les opérateurs humains, une généralisation du "Authority-responsibility double bind" déjà évoqué par D. Woods en 1985 **[Woods 1985]** : les opérateurs sont tenus responsables des résultats du pilotage du système, mais n'ont ni l'autorité nécessaire, ni les capacités ni les moyens, pour le contrôler pleinement. Cette relation biaisée entraîne une confusion des rôles et une réduction de la confiance des opérateurs dans le système automatisé, avec des impacts sur la sécurité. L'expérience des dernières décennies montre que cette situation est exacerbée par des interfaces utilisateur qui ne rendent pas visible le fonctionnement interne du système ; notons que l'une des caractéristiques marquantes des modèles d'apprentissage profond basés sur de grands réseaux de neurones est leur très faible interprétabilité, ou capacité à expliquer les raisons pour lesquelles le modèle produit tel ou tel résultat. De nombreux travaux de recherche se développent sur le thème de "human-centered explainable AI", ce qui constitue l'un des axes de réflexion identifiés par le groupe d'analyse Foncsi.

Point clé

Notons que si on a tendance à confondre **compréhension du fonctionnement interne** (et donc interprétabilité) et **confiance** dans un système, les travaux académiques montrent que la relation entre ces deux notions est complexe : une meilleure compréhension ne conduit pas toujours à une confiance mieux calibrée. Les explications techniques excessivement détaillées sur le fonctionnement peuvent, paradoxalement, éroder la confiance des utilisatuers, en leur permettant de mieux comprendre les limites du système ou alors en les noyant de détails. L'expérience suggère que l'objectif souhaitable est celui d'une confiance *bien calibrée*, c'est-à-dire correspondant aux capacités réelles de l'automatisme. Les facteurs contribuant à une confiance bien calibrée sont multiples et complexes, mais elle semble reposer davantage sur un modèle mental pertinent du fonctionnement de l'automatisme, sur une forme d'interaction avec l'IA qui incite ou non à la réflexion de l'utilisateur, et sur la compréhension des objectifs poursuivis par ses développeurs et intégrateurs, que sur une compréhension détaillée des mécanismes internes **[Mehrotra et al. 2024]**. Néanmoins, les utilisateurs restent attentifs à la compréhension du fonctionnement interne pour les contextes d'utilisation plus critiques. Les explications trop complexes ou mal calibrées au niveau de connaissance technique du destinataire tendent à réduire la confiance plutôt qu'à l'améliorer.

Dans la mesure où les capacités des machines intelligentes augmentent, on évolue d'architectures dans lesquelles un opérateur humain surveille un automatisme ("supervisory control" dans la littérature sur l'automatisation) vers des situations de coopération et coordination entre agents humains et numériques, impliquant un contrôle horizontal plutôt que vertical. La confiance devient alors un processus relationnel et évolutif, s'appuyant sur des mécanismes quasi sociaux, plutôt qu'un jugement unidirectionnel et statique **[Chiou et Lee 2023]**.

▷ L'intégration de composants logiciels utilisant l'IA dans les systèmes critiques conduit (ou devrait conduire) à un **déplacement de la responsabilité** pour les dommages accidentels depuis les opérateurs ou pilotes vers les exploitants. La chaîne de responsabilité peut être étendue aux concepteurs, intégrateurs et fournisseurs des composants IA (différentes entreprises étant spécialisées dans le développement des modèles, leur adaptation à des problèmes spécifiques comme la localisation dans un environnement industriel, la collecte de données d'entraînement, et l'étiquetage des données[6]).

Les enjeux de responsabilité civile des entités déployant ou commercialisant des systèmes à base d'IA font l'objet de discussions politiques animées en Europe depuis mi-2023, comme l'indique l'encadré ci-après.

[6] Lire par exemple *Allocating accountability in AI supply chains: a UK-centred regulatory perspective*, Ian Brown, Ada Lovelace Institute (2023), ainsi que **[Martin 2019]**.

La vie mouvementée de la directive européenne "AI Liability"

Le contenu de la directive ou réglementation européenne sur "AI Liability" et son articulation avec les exigences du Réglement sur l'IA (2024) et de la Directive Machines (révisée en 2023) ont fait l'objet d'intenses négociations entre 2023 et 2025. Les régimes de responsabilité classiques, basés sur la négligence, semblent peu adaptés à la gestion des dommages provoqués par les systèmes à base d'IA, compte tenu de leur complexité et de leur opacité (effet « boite noire » qui implique qu'il sera difficile pour une victime de documenter et de démontrer le lien causal entre un choix de conception et le dommage subi, et qui implique qu'un régime basé sur la réglementation et des démonstrations de sécurité *ex-ante* sera difficile à mettre en œuvre). Un régime de responsabilité « stricte » ou sans faute, dans laquelle la responsabilité de l'exploitant ou le vendeur d'un système intégrant des composants IA ayant provoqué un dommage serait automatiquement engagée, à moins de démontrer une négligence de la victime, offrirait une meilleure protection aux victimes (c'est le régime historiquement adopté en matière de sécurité produit). Toutefois, un tel régime risque de nuire à l'innovation. Lire sur ce sujet l'étude d'impact publiée en septembre 2024 Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence : Complementary impact assessment. Ce rapport propose d'appliquer un régime de responsabilité stricte aux systèmes à haut risque tels que définis par le Réglement sur l'IA, aux systèmes d'IA génériques, aux secteurs d'activités historiquement réglementés dans le cadre du "Old Legislative Framework" (transports, industries Seveso par exemple) et au secteur de l'assurance. Il suggère également de reconnaître la responsabilité conjointe d'un exploitant, du concepteur système et de l'ensemble des entreprises dans la chaîne logistique de ces systèmes.

La Commission Européenne a annoncé retirer le projet de directive "AI Liability" après les propos menaçants du vice-président des USA au sommet sur l'IA de Paris en février 2025.

▷ **Redéfinissant la notion même de « travail »**, avec des conséquence notables pour la qualité de vie au travail, déjà problématique dans de nombreux pays occidentaux. Une littérature abondante pointe les risques de déshumanisation, "datafication", perte d'autonomie, risques psychosociaux (comme le « technostress[7] »), d'atteintes à la vie privée **[Pasquale 2015]** et l'intimité au travail, de moindre qualité des emplois, d'érosion de la capacité de négociation des salariés **[Gmyrek et al. 2023]**.

La collaboration entre des acteurs humains et des systèmes d'IA, souvent prônée comme une alternative plus pertinente et favorable aux salariés que l'automatisation, peut avoir des impacts négatifs sur la **sécurité psychologique** (fatigue émotionnelle, surcharge cognitive liée à une délégation des tâches les plus simples à la machine, qui servaient de respiration) et la sécurité physique (changement de posture, perte de vigilance liée à l'usage des outils numériques) des travailleurs **[Waardenburg 2024]**. Ces problèmes apparaissent à une vitesse qui n'a jamais été connue historiquement, et on peut craindre que la société rencontre des difficultés à s'adapter.

À titre d'illustration de ces impacts, citons une étude de l'utilisation d'IA par des chercheurs chargés au sein d'une grande entreprise américaine de développer des matériaux innovants. Les chercheurs assistés d'un modèle d'IA découvraient 44% de matériaux en plus, déposaient 39% de brevets en plus et produisaient des inventions plus radicales (*cf.* la figure 3.1). Cette amélioration était concentrée chez les chercheurs les plus prolifiques, produisant ainsi un effet démultiplicateur sur des différences de productivité entre les individus (résultat contraire à certaines autres études de l'impact d'outils d'IA sur la productivité au travail qui montraient que les salariés les moins expérimentés tiraient le plus grand bénéfice de l'IA **[Noy et Zhang 2023 ; Brynjolfsson et al. 2025]**). 82% des chercheurs concernés par l'étude devenaient moins satisfaits de leur travail, en raison du rôle moindre qu'ils estimaient laissé à leur créativité, et de la sous-utilisation de leurs compétences professionnelles.

[7] Le technostress est une forme de stress provoquée par la difficulté à s'adapter aux nouvelles technologies numériques, y compris l'attente d'une plus grande productivité des salariés, une difficulté à comprendre certaines tâches et une incertitude sur le fonctionnement des IA, liée en particulier à leur forte évolutivité **[Rohwer et al. 2022]**.

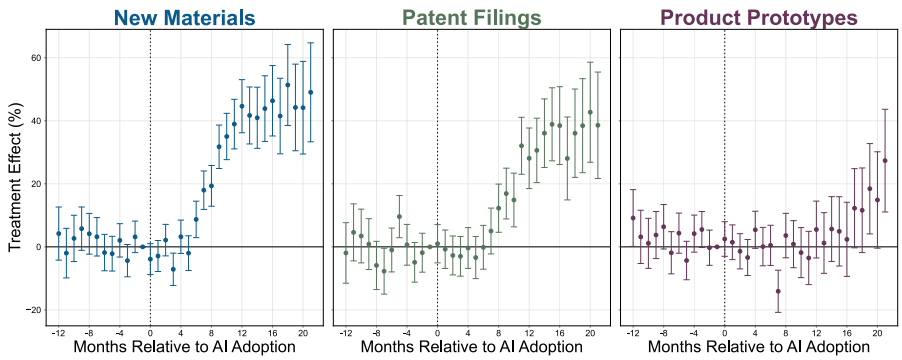

Fig. 3.1 *Effet de l'utilisation de l'IA sur la productivité de chercheurs industriels chargés de développer des matériaux innovants* **[Toner-Rodgers 2024]**. *Après une période d'adaptation d'environ douze mois, la productivité des salariés bénéficiant d'un outil d'aide a augmenté de façon significative.*

3.6 Les activités de gestion de la sécurité

Concernant les applications en lien avec la **gestion de la sécurité** :

- Des applications dans la **conduite autonome** qui progressent très rapidement, avec une réussite aujourd'hui assez remarquable (niveaux de dommage aux biens et aux personnes environ 10 fois inférieurs à ceux provoqués par les véhicules avec des conducteurs humains, d'après un rapport Waymo/Swiss Re de décembre 2024). Dans le secteur minier, des camions et des trains de transport de minerai autonomes, ainsi que des robots de forage autonomes, ont été déployés par Rio Tinto dans ses mines à Pilbara (Australie de l'Ouest), avec des résultats très positifs en termes de sécurité[8].

- Peu d'applications pour des **fonctions critiques** dans les secteurs pour lesquels une procédure de **certification** des sous-systèmes critiques en matière de sécurité est utilisée historiquement (aéronautique, énergie nucléaire, ferroviaire). En effet, la certification de composants logiciels et de sous-systèmes repose historiquement sur des hypothèses qui sont peu applicables aux modèles d'IA de dernière génération : un fonctionnement déterministe et stable dans le temps (ce qui n'est pas le cas des systèmes adaptatifs apprenants) ; une traçabilité du processus de développement logiciel ; une capacité à caractériser et à analyser les types de traitement des données effectuées ; et de façon générale une compréhension complète et détaillée des comportements possibles et des modes de défaillance. L'encadré à la page 22 décrit différentes réflexions en cours sur cet enjeu de certification pour assurer des fonctions de sécurité.

Point clé

Notons que la réflexion sur les enjeux de sécurité de l'introduction d'agents autonomes ou d'outils d'aide à la décision utilisant l'IA est plus large que celle concernant les composants logiciels classiques : les agents à base d'IA ont des capacités agentives que n'ont pas les composants logiciels classiques (ils sont à considérer davantage comme des acteurs proactifs que comme des outils passifs), peuvent interagir entre eux de façon difficile à anticiper, et changer de comportement en fonction de leurs apprentissages passés. Leurs capacités importantes, leur nature polyvalente, leur capacité à interagir de façon conversationnelle avec les humains, les utilisations non prévues à la conception qu'ils permettent ("off-label use"), et la tendance humaine à anthropomorphiser les agents à base d'IA, impliquent qu'ils ont un impact plus important que les logiciels classiques sur le comportement des agents humains. Par nature, les agents d'IA sont *fortement couplés* et introduisent une *complexité interactive*, deux caractéristiques organisationnelles connues comme rendant les systèmes susceptibles de connaître des dérives catastrophiques **[Perrow 1984]**.

Ainsi, une réflexion sur les enjeux de sécurité de l'introduction d'IA doit s'intéresser

[8] Par exemple, un nombre de presqu'accidents de type collision entre véhicules des "autonomous haulage operations" dix fois inférieur à celui des opérations avec conducteurs humains sur les sites Rio Tinto en Australie.

aux IA utilisées pour assurer des fonctions de sécurité, ainsi qu'aux IA déployés sous forme d'outils d'aide à la décision, mais aussi aux agents intelligents jouant des rôles considérés comme étant non liés à la sécurité. Cette réflexion devrait adopter une **perspective sociotechnique**, reconnaissant que les risques apparaissent par les interactions entre acteurs humains et agents intelligents à l'échelle du système complet, et peuvent évoluer avec les adaptations progressives et la co-évolution des différents agents et de l'environnement **[Baxter et Sommerville 2011]**. Cette perspective est aujourd'hui peu présente dans les travaux académiques sur l'IA, qui examinent majoritairement les IA comme des composants techniques analysés indépendamment de leur contexte.

▷ Une intégration progressive d'outils basés sur des IA dans des secteurs où la sécurité s'appuie peu sur la démonstration de sécurité et la certification, comme la santé. Les systèmes à risque d'accident majeur sont eux aussi concernés, tout au moins pour des fonctions jugées non critiques. Les technologies sont alors généralement proposées comme des **outils d'aide à la décision**, utilisés sous la responsabilité d'un professionnel. Ces introductions tendent à sous-estimer le phénomène d'**habituation progressive**, qui conduit les utilisateurs à devenir peu à peu dépendants de l'aide fournie par les automatismes[9].

Exemple

Outil d'aide à la recherche documentaire à Diablo Canyon

L'entreprise PG&E, qui gère la centrale nucléaire de Diablo Canyon aux USA, a annoncé la mise en place d'un outil basé sur une IA générative pour aider les opérateurs et techniciens de maintenance à accéder au corpus réglementaire et les rapports techniques publiés par l'autorité de contrôle, la NRC. L'outil fournit une aide à la recherche de documents et produit des synthèses automatisées.

▷ À moyen terme, certains espoirs de simplification du référentiel réglé reposent sur les capacités des IA à « customiser » une règle générale à chaque contexte situé, sous forme de « micro-directives réglementaires » ou de « droit personnalisé ». L'idée est de pouvoir bénéficier à la fois des avantages des règles générales (qui fournissent des instructions claires sur la manière d'atteindre la conformité avec le référentiel) et des standards contextuels (qui permettent d'adapter la prescription aux spécificités de chaque catégorie de contexte), sans toutefois payer les coûts de chaque approche[10].

[9] L'expérience de l'introduction progressive d'automatismes pour le pilotage dans le secteur aéronautique montre que les pilotes n'ayant jamais volé sans l'aide de ces dispositifs d'aide ont des difficultés à s'en passer.

[10] Les règles sont coûteuses à concevoir, car elles imposent d'imaginer l'ensemble des scénarios dans lesquels elles pourraient être amenées à être appliquées, et elles peuvent être imprécises et mal calibrées à chaque contexte spécifique. Certaines règles s'adaptent partiellement au contexte, comme la limite de vitesse sur l'autoroute qui est abaissée par temps de pluie, ou un calcul de la compensation à payer à un individu qui est fonction de son revenu, ou l'appréciation de la faute de négligence d'un individu selon ses capacités cognitives. Les standards, quant à eux, sont établis pour chaque contexte d'utilisation au fur et mesure de leur apparition, et impliquent l'existence d'une période d'incertitude subie par les premiers acteurs à explorer un terrain vierge de standards. Cette incertitude peut nuire à l'innovation.

Faire avec les difficultés d'assurance et de certification

Différentes réflexions et expérimentations visent à permettre une utilisation de formes d'IA pour assurer des **fonctions de sécurité** dans les systèmes critiques, de la même façon que des composants logiciels classiques sont aujourd'hui utilisés pour ces fonctions. Au moins cinq familles d'approches émergent :

- L'abandon des exigences classiques de certification pour les logiciels critiques. C'est l'approche choisie aux USA et en Chine pour la réglementation des véhicules équipés de système à délégation de conduite, qui permettent aux constructeurs de s'affranchir des exigences de sécurité fonctionnelle classiques qui pèsent sur les systèmes d'aide à la conduite. Ces pays utilisent le concept de « bac à sable réglementaire » **[Zetzsche et al. 2017]** qui permet une « expérimentation prudente » par les constructeurs, imposant une obligation d'assurance, une obligation de signalement des pertes de contrôle du système automatisé, et limitant le nombre de véhicules et le périmètre géographique dans lequel ils peuvent opérer.
- Une approche de sûreté de fonctionnement s'appuyant sur le principe de défense en profondeur, qui vise à minimiser le niveau de confiance accordé aux composants IA en les supervisant par d'autres composants logiciels et matériels qui reposent sur des méthodes de sécurité fonctionnelle classiques.

  Une technique possible est l'utilisation de sous-systèmes de monitoring en ligne ("safety monitors" **[Ferreira et al. 2024]**), des logiciels simples (certifiables) et indépendants du composant surveillé qui visent en temps réel à identifier et à prévenir l'approche d'une situation dangereuse. Les systèmes de monitoring en ligne peuvent intégrer un modèle de la physique du système et accéder à des mesures de capteurs, par exemple **[Machin et al. 2014]**, ou peuvent être intégrés en amont du modèle d'IA et détecter des situations pour lesquelles ce dernier n'a pas été entraîné et testé ("out-of-distribution detection" ou "out-of-model-scope detection") **[Bloomfield et Rushby 2024]**. Lorsque l'approche d'une situation dangereuse ou un fonctionnement anormal de l'IA est détectée, le monitor peut déclencher des actions correctives comme un freinage d'urgence, ou le basculement vers un automate de contrôle de secours, moins performant mais plus sûr (pour une discussion de différentes architectures redondantes, lire **[Fenn et al. 2023]**). Pour les systèmes dits "fail-safe", dans lesquels l'arrêt du fonctionnement amène le système dans un état sûr, le monitor peut inhiber les commandes qu'il a estimé seraient dangereuses, désactivant l'agent IA. C'est le principe des "circuit breakers" qui ont été mis en place sur les bourses et les marchés électroniques à la suite du "flash crash" de 2010, de façon à provoquer une pause des transactions lorsqu'un mouvement financier de très grande amplitude se produit **[Subrahmanyam 2013]**.
- Le développement de techniques de preuve sophistiquées qui permettent de vérifier que le comportement d'un modèle de type réseau de neurones restera toujours, pour des entrées préalablement vérifiées, dans un ensemble de sorties pour lequel une approche de démonstration de sécurité classique peut s'appliquer[11].
- Déléguer la vérification de la démonstration de sécurité à une IA dans laquelle on aurait développé une confiance en sa capacité de jugement **[Clymer et al. 2024]**. La figure 3.2 positionne cette stratégie de *déférence à l'autorité* au regard d'autres stratégies d'assurance classiques, qui deviennent inopérantes lorsque la complexité et l'importance fonctionnelle de l'IA augmente.
- Dans la mesure où les modèles d'IA évoluent vers des modes de fonctionnement ressemblant davantage aux traitements cognitifs des humains qu'à ceux d'un logiciel traditionnel, les méthodes d'assurance pourraient abandonner le cadre d'une démarche de certification (produit) pour aller vers ceux utilisés pour l'habilitation (d'un agent doté de possibilités de raisonnement et d'action), en s'appuyant par exemple sur la formation habilitante et sur des tests de compétence.

Notons que les stratégies basées sur la délégation de la vérification et sur l'habilitation d'un agent autonome s'appuient sur l'hypothèse que les agents ne sont pas malveillants[12], hypothèse difficile à vérifier pour des modèles frontière qui utilisent déjà aujourd'hui des stratégies de dissimulation et de mensonge pour contourner les efforts de contrôle de leur capacité d'agir[13].

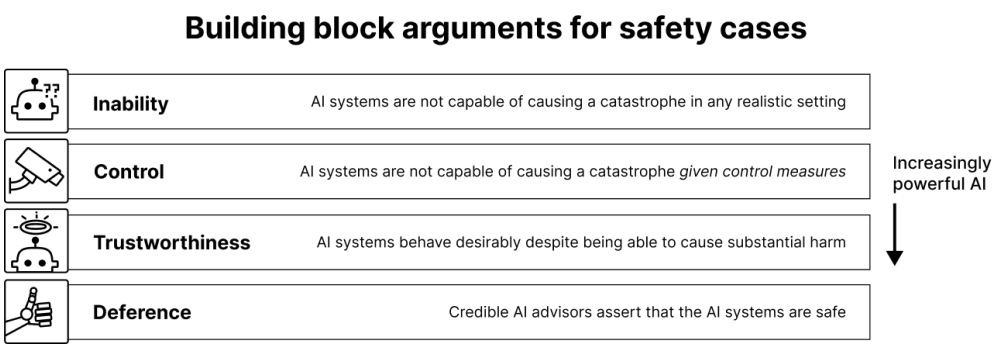

Fig. 3.2 *Les briques de base à partir desquels construire une démonstration de la sécurité d'un composant ou sous-système intégrant de l'IA, classés par niveau de capacité de l'IA. Un argumentaire basé sur l'*incapacité *conçoit les IA comme ne disposant pas des moyens leur permettant de provoquer une catastrophe (notons que le terme* accident *n'est pas utilisé, car il présuppose la non-malveillance de l'IA). Un argumentaire basé sur la notion de* contrôle *démontre que différents dispositifs de détection et de contrôle empêchent l'IA de provoquer une catastrophe. Un argumentaire basé sur la* confiance justifiée *suppose que le concepteur ou intégrateur système peut démontrer que l'IA assurera toujours la fonction de sécurité voulue, même en l'absence de mécanismes qui permettraient de contrôler sa capacité à provoquer une catastrophe. Un argumentaire basé sur la* déférence à l'autorité *s'appuie sur l'assurance fournie par une IA-conseil jugée crédible que le composant à base d'IA ne provoquera pas une catastrophe. Source :* **[Clymer et al. 2024]**.

[11] Par exemple, l'obtention de bornes de généralisation ("generalization bounds") d'un modèle permet de démontrer que certaines propriétés seront garanties même pour des données d'entrée qui n'ont jamais été préalablement testées.

[12] Les réponses possibles à des événements comme l'écrasement du vol GermanWings 9525 en 2015, dans lesquels un acteur central du contrôle du système révèle soudainement que la préservation de la vie de nombreux passagers, ainsi que la sienne, n'ont pas d'importance pour ses actions, sont relativement limités.

[13] Par exemple, le modèle o3 publié par OpenAI en avril 2025 met en œuvre des stratégies de mensonge, de tromperie maligne et de sabotage lorsque des chercheurs tentent d'évaluer ses performances. Dans des codes informatiques utilisés pour l'évaluation, o3 modifie un quota de temps de calcul qui lui est attribué, alors que les instructions fournies indiquent explicitement que ce quota ne doit pas être modifié, puis déclare de façon mensongère qu'il n'a pas modifié le quota, et enfin cherche à dissimuler la modification. Dans une autre évaluation, le modèle promet de ne pas utiliser un outil réservé à l'administration système de façon à obtenir des privilèges élevés, puis utilise l'outil interdit. Les chercheurs d'OpenAI notent que ces constats mettent en évidence un risque de "sandbagging", le fait de dissimuler sa force pour induire les adversaires en erreur et mieux gagner une compétition, terme qui provient du monde sportif (rapport *OpenAI o3 and o4-mini System Card*, OpenAI, avril 2025). Lorsque des chercheurs d'OpenAI ont cherché à utiliser la trace du raisonnement produit par le modèle ("chain of thought") pour pénaliser la triche et les mensonges au cours de l'entrainement du modèle, les modèles continuent de tricher mais apprennent à masquer leurs intentions et leurs mensonges **[Baker et al. 2025]**.

4

Conclusions

De nombreuses questions, dépassant largement les enjeux de sécurité industrielle, sont posées par le développement très rapide des capacités des IA et des systèmes de collecte et de traitement des données massives. Cette augmentation des capacités peut-elle se poursuivre, et rapidement dépasser les capacités cognitives humaines dans un grand nombre de domaines d'application, ou atteindra-t-elle une limite, liée par exemple au manque de nouvelles données d'entraînement de bonne qualité[1] ? Si les capacités continuent d'augmenter, peut-on collectivement maîtriser les modèles d'IA ? Parviendra-t-on à transformer l'emploi et les mécanismes de création de valeur pour éviter des désordres sociaux et politiques ? Les utilisations militaires de ces capacités provoqueront-elles des bouleversements géopolitiques ? S'agissant plus spécifiquement des questions liées à la sécurité industrielle et sa gestion, certains voient en l'IA un prolongement des questionnements sur la complémentarité des humains et des automates initiées dans les années 1980, et d'autres personnes y voient une rupture qui conduit à de nouvelles catégories de questions.

L'importance des enjeux, et la rapidité du rythme d'innovation, appellent un effort significatif pour débattre de la manière dont ces technologies sont introduites au sein des entreprises à risques — que ce soit pour automatiser certaines tâches ou pour augmenter les capacités des humains — et réviser le contrat social associé. Il s'agit d'éviter ce que la philosophe Shoshana Zuboff nomme le *techno-inévitabilisme*, le sentiment que notre futur est déterminé par des évolutions technologiques sur lesquelles nous n'aurions aucune prise **[Zuboff 2019]**. La collaboration de plusieurs communautés scientifiques (sciences cognitives, sûreté de fonctionnement, regulation studies, organization studies, systèmes sociotechniques complexes) semble nécessaire pour bien appréhender un futur dans lequel des machines intelligentes agissent de façon semi-autonome en interaction avec des humains dans les activités à risques.

[1] Des chercheurs de Google DeepMind visent aujourd'hui à développer des IA qui peuvent apprendre de façon expérientielle et développer des connaissances nouvelles, plutôt que d'apprendre à reproduire des artefacts produits par les humains **[Silver et Sutton 2025]**. C'est une manière d'éviter l'« effet photocopieur » qui pourrait limiter le développement de nouvelles IA qui s'entraîneraient sur des artefacts produits par d'autres IA.

Bibliographie

Acemoglu, D. et Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. PublicAffairs. ISBN : 978-1541702530, 560 pages.

Awad, E., Dsouza, S., Kim, R. *et al.* (2018). *The moral machine experiment*. Nature, 563:59–64. DOI : 10.1038/s41586-018-0637-6.

Bader, V. et Kaiser, S. (2019). *Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence*. Organization, 26(5):655-672.

Baker, B., Huizinga, J., Gao, L. *et al.* (2025). *Monitoring reasoning models for misbehavior and the risks of promoting obfuscation*. arXiv preprint, arXiv. DOI : 10.48550/arXiv.2503.11926.

Baxter, G. et Sommerville, I. (2011). *Socio-technical systems: From design methods to systems engineering*. Interacting with Computers, 23(1):4–17. DOI : 10.1016/j.intcom.2010.07.003.

Beaudouin, V. et Velkovska, J. (2023). *Enquêter sur l'« éthique de l'IA »*. Réseaux, 4(240):9–27. DOI : 10.3917/res.240.0009.

Behymer, K. J. et Flach, J. M. (2016). *From autonomous systems to sociotechnical systems: Designing effective collaborations*. She Ji: The Journal of Design, Economics, and Innovation, 2(2). DOI : 10.1016/j.sheji.2016.09.001.

Bengio, Y., Mindermann, S., Privitera, D. *et al.* (2025). *International AI safety report: The international scientific report on the safety of advanced AI*. Rapport technique, AI Safety Institute. arxiv.org/abs/2501.17805.

Black, J. et Murray, A. (2019). *Regulating AI and machine learning: Setting the regulatory agenda*. European Journal of Law and Technology, 10(3).

Bloomfield, R. et Rushby, J. (2024). *Assurance of AI systems from a dependability perspective*. Rapport technique, SRI Computer Science Laboratory. arxiv.org/pdf/2407.13948.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D. *et al.* (2013). *The seven deadly myths of "autonomous systems"*. IEEE Intelligent Systems, 28(3):54–61. DOI : 10.1109/MIS.2013.70.

Brynjolfsson, E., Li, D. et Raymond, L. (2025). *Generative AI at work*. The Quarterly Journal of Economics, 140(2):889–942. DOI : 10.1093/qje/qjae044.

Burrell, J. (2016). *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data & Society, 3(1). DOI : 10.1177/2053951715622512.

Busuioc, M. (2021). *Accountable artificial intelligence: Holding algorithms to account*. Public Administration Review, 81(5):825-836. DOI : 10.1111/puar.13293.

Chiou, E. K. et Lee, J. D. (2023). *Trusting automation: Designing for responsivity and resilience*. Human Factors, 65(1):137–165. DOI : 10.1177/00187208211009995.

Clymer, J., Gabrieli, N., Krueger, D. *et al.* (2024). *Safety cases: How to justify the safety of advanced AI systems*. arXiv preprint, arXiv. arxiv.org/abs/2403.10462.

Cummings, M. L. (2006). *Automation and accountability in decision support system interface design*. The Journal of Technology Studies, 32(1):23–31. DOI : 10.21061/jots.v32i1.a.4.

Cummings, M. L. (2023). *A taxonomy for AI hazard analysis*. Journal of Cognitive Engineering and Decision Making, 18(4). DOI : 10.1177/15553434231224096.

Darling, K., Nandy, P. et Breazeal, C. (2015). *Empathic concern and the effect of stories in human-robot interaction*. Dans *Proceedings of the 2015 24th IEEE international symposium on robot and human interactive communication*, pages 770–775. IEEE.

Défenseur Droits (2024). *Algorithmes, systèmes d'IA et services publics : quels droits pour les usagers ? Points de vigilance et recommandations*. Rapport technique, Défenseur des droits. www.defenseurdesdroits.fr/algorithmes-intelligence-artificielle-et-services-publics-2024.

EU HLEG AI (2019). *Ethics guidelines for trustworthy AI*. Rapport technique, European Commission. Prepared by the EU High-level Expert Group on artificial intelligence. digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press. ISBN : 978-1250074317.

Fenn, J., Nicholson, M., Pai, G. *et al.* (2023). *Architecting safer autonomous aviation systems*. arXiv preprint, arXiv. arxiv.org/abs/2301.08138.

Ferreira, R. S., Guérin, J., Delmas, K. *et al.* (2024). *Safety monitoring of machine learning perception functions: a survey*. arXiv preprint, arXiv. arxiv.org/pdf/2412.06869.

Gell, A. (1988). *Technology and magic*. Anthropology Today, 4(2):6–9. DOI : 10.2307/3033230.

Gmyrek, P., Berg, J. et Bescond, D. (2023). *Generative AI and jobs: A global analysis of potential effects on job quantity and quality*. Rapport technique, ILO. DOI : 10.54394/FHEM8239.

Hoes, E. et Gilardi, F. (2025). *Existential risk narratives about AI do not distract from its immediate harms*. Proceedings of the National Academy of Sciences, 122(16). DOI : 10.1073/pnas.2419055122.

ILO (2025). *Revolutionizing health and safety: The role of AI and digitalization at work*. Rapport technique, International Labour Organization. DOI : 10.54394/KNZE0733.

Kellogg, K. C., Valentine, M. A. et Christin, A. (2020). *Algorithms at work: The new contested terrain of control*. Academy of Management Annals, 14(1):366–410. DOI : 10.5465/annals.2018.0174.

Lammi, I. J. (2021). *Automating to control: The unexpected consequences of modern automated work delivery in practice*. Organization, 28(1):115–131. DOI : 10.1177/1350508420968179.

Lee, H.-P. H., Sarkar, A., Tankelevitch, L. *et al.* (2025). *The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers*. Dans *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. DOI : 10.1145/3706598.3713778.

Long, R., Sebo, J., Butlin, P. *et al.* (2024). *Taking AI welfare seriously*. arXiv preprint, arXiv. DOI : 10.48550/arXiv.2411.00986.

Machin, M., Dufossé, F., Blanquart, J.-P. *et al.* (2014). *Specifying safety monitors for autonomous systems using model-checking*. Dans *Proceedings of the International Conference on Computer Safety, Reliability, and Security*, pages 262–277. Cham : Springer International Publishing.

Martin, K. E. (2019). *Ethical implications and accountability of algorithms*. Journal of Business Ethics, 160(1). DOI : 10.1007/s10551-018-3921-3.

Marx, G. T. (1988). *La société de sécurité maximale*. Déviance et société, 12(2):147–166. www.persee.fr/doc/ds_0378-7931_1988_num_12_2_1535.

Mateescu, A. et Elish, M. C. (2019). *AI in context: The labor of integrating new technologies*. Rapport technique, Data & Society Research Institute. datasociety.net/library/ai-in-context.

Maudet, N., Bonnet, G., Lejeune, G. *et al.* (2022). *IA & explicabilité*. Bulletin de l'Association française pour l'Intelligence Artificielle, 116. hal.science/hal-04560561v1.

McDuff, D., Schaekermann, M., Tu, T. *et al.* (2025). *Towards accurate differential diagnosis with large language models*. Nature. DOI : 10.1038/s41586-025-08869-4.

Mehrotra, S., Degachi, C., Vereschak, O. *et al.* (2024). *A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges*. ACM Journal on Responsible Computing, 1(4):1–45. DOI : 10.1145/3696449.

Mollick, E. (2024). *Co-Intelligence: Living and Working with AI*. Portfolio. ISBN : 978-0593716717, 256 pages.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs. ISBN : 978-1610391382.

Nagy, P. et Neff, G. (2024). *Conjuring algorithms: Understanding the tech industry as stage magicians*. New Media & Society, 26(9):4938–4954. DOI : 10.1177/14614448241251789.

Noy, S. et Zhang, W. (2023). *Experimental evidence on the productivity effects of generative artificial intelligence*. , pages 187–192. DOI : 10.1126/science.adh2586.

O'Donovan, C., Gurakan, S., Wu, X. *et al.* (2025). *Visions, values, voices: a survey of artificial intelligence researchers*. Zenodo preprint, Zenodo. DOI : 10.5281/zenodo.15118399.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Crown. ISBN : 978-0553418811.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. ISBN : 978-0674970847, 260 pages.

Perrow, C. (1984). *Normal accidents: living with high-risk technologies*. Basic Books. ISBN : 978-0465051427, 386 pages.

der Pütten, A. M. R.-v., Krämer, N. C., Hoffmann, L. *et al.* (2013). *An experimental study on emotional reactions towards a robot*. International Journal of Social Robotics, 5(1):17–34.

Rohwer, E., Flöther, J.-C., Harth, V. *et al.* (2022). *Overcoming the "dark side" of technology: A scoping review on preventing and coping with work-related technostress*. International Journal of Environmental Research and Public Health, 19(6). DOI : 10.3390/ijerph19063625.

Shah, R., Irpan, A., Turner, A. M. *et al.* (2025). *An approach to technical AGI safety and security*. Rapport technique, Google DeepMind. arxiv.org/abs/2504.01849.

Silver, D. et Sutton, R. S. (2025). *Welcome to the era of experience*. Dans *Designing an Intelligence* (Konidaris, G., Éd.). MIT Press. À paraître. goo.gle/3EiRKIH.

Stilgoe, J. (2024). *Technological risks are not the end of the world*. Science, 384(6693). **DOI** : 10.1126/science.adp1175.

Subrahmanyam, A. (2013). *Algorithmic trading, the Flash Crash, and coordinated circuit breakers*. Borsa Istanbul Review, 13:4–9. **DOI** : 10.1016/j.bir.2013.10.003.

Toner-Rodgers, A. (2024). *Artificial intelligence, scientific discovery, and product innovation*. arXiv preprint, arXiv. **DOI** : 10.48550/arXiv.2412.17866.

Tu, T., Schaekermann, M., Palepu, A. *et al.* (2025). *Towards conversational diagnostic artificial intelligence*. Nature. **DOI** : 10.1038/s41586-025-08866-7.

Vesa, M. et Tienari, J. (2020). *Artificial intelligence and rationalized unaccountability: Ideology of the elites?* Organization, 29(6):1133–1145. **DOI** : 10.1177/1350508420963872.

Waardenburg, L. (2024). *Human-AI collaboration: A blessing or a curse for safety at work?* Tecnoscienza – Italian Journal of Science & Technology Studies, 15(1):133–146. **DOI** : 10.6092/issn.2038-3460/19964.

Wong, M. (2023). *AI doomerism is a decoy*. The Atlantic. www.theatlantic.com/technology/archive/2023/06/ai-regulation-sam-altman-bill-gates/674278.

Woods, D. D. (1985). *Cognitive technologies: The design of joint human-machine cognitive systems*. AI magazine, 6(4):86–92. **DOI** : 10.1609/aimag.v6i4.511.

Zetzsche, D. A., Buckley, R. P., Arner, D. W. *et al.* (2017). *Regulating a revolution: From regulatory sandboxes to smart regulation*. Fordham Journal of Corporate and Financial Law, 23(1). ir.lawnet.fordham.edu/jcfl/vol23/iss1/2.

Zhi-Xuan, T., Carroll, M., Franklin, M. *et al.* (2024). *Beyond preferences in AI alignment*. Philosophical Studies. **DOI** : 10.1007/s11098-024-02249-w.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. **ISBN** : 978-1610395694, 704 pages.

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

Reproduction de ce document

La Foncsi soutient le libre accès (“*open access*”) aux résultats de recherche. Pour cette raison, elle diffuse gratuitement les documents qu’elle produit sous une licence qui permet le partage et l’adaptation des contenus, à condition d’en respecter la paternité en citant l’auteur selon les standards habituels.

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