Rapport à Monsieur le Premier ministre

Mission parlementaire sur l'alignement des systèmes d'intelligence artificielle (SIA)

Pour une filière française -et européenne- de

l'IA alignée

articulant robustesse technique, exigences démocratiques, maîtrise des dépendances critiques et résilience collective.

Vanina PAOLI-GAGIN,

Sénateur de l'Aube

Paris, le 29 juillet 2026

Résumé exécutif

Face à des systèmes d'intelligence artificielle (SIA) plus autonomes, plus opaques et plus intégrés à des infrastructures critiques, qui conditionnent une part croissante de l'ensemble des activités humaines, l'Europe ne peut se contenter de réguler *a posteriori*, ni de dépendre d'évaluations produites ailleurs. **Le rapport ouvre une autre perspective à la France : retrouver une part de souveraineté numérique en faisant de l'alignement des SIA une filière scientifique industrielle et démocratique.**

L'alignement est une notion faîtière qui abrite et fait coexister trois exigences complémentaires :

- ✓ **technique** : rendre les systèmes plus robustes, fiables, contrôlables, évaluables et vérifiables ;
- ✓ **normative et éthique** : garantir que les systèmes respectent les finalités humaines, les droits fondamentaux, les principes démocratiques et les limites fixées par la société ;
- ✓ **systémique** : maîtriser les conditions de développement, de déploiement, de supervision et d'usage des systèmes d'IA.

La France doit agir pour consolider la gouvernance, renforcer la recherche en interprétabilité, développer des standards de preuve et structurer un marché européen solvable de l'IA alignée. La commande publique et privée, les infrastructures de calcul, l'énergie, les données et la normalisation sont des leviers stratégiques à actionner sans délai **pour que l'alignement devienne aux SIA ce que la sûreté a été à l'aéronautique : une exigence transformée en standard de puissance industrielle exporté dans le monde entier.** L'enjeu est économique, mais surtout démocratique et existentiel ; il implique de définir ce que l'on consent à déléguer aux machines, dans quelles conditions et au nom de quelles valeurs.

Executive summary

Faced with systems that are more autonomous, more opaque, and more deeply integrated into critical digital infrastructures, Europe cannot simply regulate after the fact or rely on assessments produced elsewhere. The report proposes a shift in perspective at France's initiative: recovering a part of our digital sovereignty by making AI alignment a scientific, industrial, as well as a democratic endeavor.

Alignment is an overarching concept that encompasses and brings together three complementary requirements:

- ✓ technical: making systems more robust, reliable, controllable, assessable, and verifiable;
- ✓ normative and ethical: ensuring that systems respect human purposes, fundamental rights, democratic principles, and the limits set by society;
- ✓ systemic: managing the conditions for development, deployment, supervision, and use.

France must take action to consolidate governance, strengthen research on interpretability, develop standards of evidence, and establish a viable European market for aligned AI. Public procurement and private demand, computing infrastructure, energy, data, and standardization are strategic levers that must be activated without delay so that alignment becomes what safety has been for the aerospace industry: a requirement transformed into an industrial standard exported worldwide. The stakes are economic, but above all democratic and existential, in that they involve defining what we consent to delegate to machines, under what conditions, and in the name of what values.

Feuille de route : dix priorités à court terme pour l'exécutif

Priorité n° 1 : Porter une doctrine française et européenne commune de l'alignement
Objectif — Stabiliser une définition partagée de l'alignement et traduire cette doctrine en exigences proportionnées à la performance des systèmes, à la criticité de leurs usages et au caractère réversible ou irréversible de leur diffusion.

Priorité n° 2 : Placer l'humain, l'éthique et la littératie au cœur des usages
Objectif — Faire de la participation des utilisateurs, de la délibération éthique, de la supervision humaine et de la compréhension des systèmes des composantes constitutives de l'alignement.

Priorité n° 3 : Faire de l'interprétabilité et de la contrôlabilité une priorité de recherche
Objectif — Rééquilibrer l'effort de recherche consacré à l'augmentation des capacités des modèles en investissant dans la compréhension, la contrôlabilité et la vérification de leur fonctionnement.

Priorité n° 4 : Construire une capacité indépendante d'évaluation et de vérification
Objectif — Donner aux autorités publiques et aux chercheurs les moyens d'accéder aux modèles, d'évaluer leurs capacités et de vérifier les garanties avancées avant et après leur déploiement.

Priorité n° 5 : Bâtir une infrastructure de preuve, de traçabilité et d'auditabilité
Objectif — Passer d'une conformité essentiellement déclarative à des garanties fondées sur des faits vérifiables, des traces opposables et l'observation du comportement réel des systèmes.

Priorité n° 6 : Doter la France d'une gouvernance opérationnelle de l'alignement
Objectif — Consolider les institutions existantes, clarifier leurs responsabilités et leur donner les moyens humains, scientifiques et techniques nécessaires pour agir au rythme des capacités de l'IA.

Priorité n° 7 : Transformer la conformité en accompagnement et en avantage compétitif
Objectif — Donner aux entreprises des parcours lisibles de mise en conformité, développer l'offre d'audit et de certification et faire des exigences européennes un facteur de confiance, d'adoption et de différenciation.

Priorité n° 8 : Faire de l'État et de la commande un moteur de l'IA alignée
Objectif — Utiliser les achats, les infrastructures et les usages publics pour structurer la demande, diffuser des exigences vérifiables et éviter la duplication de solutions coûteuses ou insuffisamment maîtrisées.

Priorité n° 9 : Structurer une base industrielle et technologique de l'IA alignée
Objectif — Concentrer l'effort public et privé sur les briques rares et créatrices de valeur et maîtriser les conditions d'accès aux ressources critiques.

Priorité n° 10 : Préparer la résilience et porter une diplomatie de l'alignement
Objectif — Anticiper la défaillance des garanties, préparer la continuité des fonctions essentielles, organiser la reprise de maîtrise et porter une coopération internationale adaptée.

Ces priorités regroupent les recommandations à court et moyen termes, parmi les 40 formulées dans le rapport. Leurs leviers d'action sont détaillés dans la Partie III.

Contexte : Servitude sur servitude ne vaut

> « ***Pour que les hommes, tant qu'ils sont des hommes, se laissent assujettir, il faut de deux choses l'une : ou qu'ils y soient contraints, ou qu'ils soient trompés.*** »[1]
>
> Étienne de la Boétie

Élue de la Nation attachée à la liberté – notre liberté chérie ! –, je ne nous souhaite ni la contrainte, ni la tromperie : si nous n'y prenons garde, elles pourraient être, s'agissant des systèmes d'intelligence artificielle (SIA), le double visage d'un même Janus.

Parce que l'IA, c'est politique ; impression que j'avais ressentie en 2016, lors de mon premier sommet international sur le sujet à Hangzhou, en Chine. Depuis, j'ai observé, comme tout un chacun, l'accélération prodigieuse du potentiel des SIA déployés dans le monde entier. Cette accélération a sans cesse renforcé mon impression première, et partant augmenté les doutes sur les effets induits pour notre société : comment s'assurer que cette rupture technologique demeure alignée avec les valeurs qui en constituent le ciment premier ?

Pour faire émerger des réponses sérieuses, j'ai donc organisé, le 27 octobre 2025, un colloque au Palais du Luxembourg sur l'alignement des SIA. Inspirée par la lecture de l'ouvrage *Hyperarme* de Flavien Chervet[2], j'ai ainsi réuni des spécialistes français et européens – chercheurs, industriels, investisseurs, experts, décideurs publics… – dont la Ministre déléguée à l'IA et au Numérique, Mme Anne Le Hénanff. Il s'agissait de mesurer collectivement la réalité de ce que je percevais déjà alors comme une formidable opportunité : l'alignement.

Inquiète de nos décrochages et de nos dépendances systémiques, je me lassais de nous voir singer encore, avec plus ou moins de bonheur et toujours un train de retard, les États-Unis. Mon intuition était que, face à la déferlante sur le territoire européen des SIA venus d'ailleurs, nous pourrions trouver une place —ou plutôt : créer une place qui serait la nôtre par ce qu'elle dirait de nous.

Notre place, pour être légitime, doit résonner avec notre histoire, notre culture et notre conception de la liberté et de la démocratie ; pour être incontestable à la lumière de notre droit et de notre patrimoine européen, elle doit promouvoir l'humanisme, qui défend l'homme hybridé[3], plutôt que le transhumanisme. Aussi notre place doit-elle nous ériger en défenseurs, en garants et en fournisseurs d'un alignement technique, éthique et systémique des SIA. C'était déjà le sens de ma proposition de résolution européenne, déposée au Sénat en octobre 2025[4].

L'alignement nous invite, il me semble, à penser l'IA comme un écosystème vivant, dont la stabilité dépend plus de la qualité des interactions que d'un ordre rigide et centralisé. Alors que cette « forêt intelligente » menace de brûler, j'ai eu à cœur de tisser un « fil vert » au long de ce rapport : charge à nous d'en faire la (photo)synthèse ! Cet alignement, nous nous le devons, et, surtout, nous le devons à la jeune génération qui constate ô combien que « *se dessaisir du courage est un mauvais calcul puisque le coût de la lâcheté est bien supérieur à celui du courage*. »[5]

Osons donc le courage ! et puis nous verrons bien…

---

[1] Etienne de la Boétie, *Discours de la servitude volontaire*, 1576

[2] Flavien Chervet, *Hyperarme : Anatomie d'une intranquilité planétaire*, Nullius in Verba, 2025

[3] Gabrielle Halpern, *Tous centaures ! Eloge de l'hybridation*, PUF, 2020

[4] Sénat - Proposition de résolution européenne n°53 de Vanina Paoli-Gagin du 22 octobre 2025, PPRE

[5] Cynthia Fleury, Université de Genève, Campus n°102, 2012, p.28

Remerciements

Alors, merci Monsieur le Premier ministre de m'avoir accordé votre confiance.

Merci à ma fidèle équipe parlementaire de mustangs (sans harnais !) avec une mention spéciale à Léo Courvasier (pour son endurance et sa fougue). Merci à Marc-Antoine Authier (pour sa structuration de notre colloque, son art d'agrégateur de talents et de communicant) et à Jean-Michel Adélaïde (pour sa capacité à recadrer mon imagination, pour ses compétences d'ingénieur qui ont évité quelques hallucinations, pour son travail de synthèse et de mise en opérationnalité de nos recommandations). Merci aussi à Joud Assouad, stagiaire ENS qui a contribué à notre travail de réflexion.

Merci à l'équipe *pro-bono* resserrée de très grande qualité, qui a bien voulu donner de son temps sans compter, livrer des clés de compréhension et contribuer activement à la réflexion :

- Laurence Devillers, professeure à l'université Paris-Sorbonne, directrice de recherche au Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) ;
- Florent Kirchner, directeur du pôle souveraineté numérique du Secrétariat général pour l'investissement (SGPI) ;
- Tom David, cofondateur et président-directeur général de GPAI Policy Lab ;
- Chloé Touzet et Alice Teilhard de Chardin, respectivement *head of policy* et *policy associate* à SaferAI ;
- Flavien Chervet, prospectiviste en IA et auteur.

Merci à Clément Duport pour la mise à disposition de Pocket Consultant, solution de génération à enrichissement contextuel (RAG), développée par PolysemIQ et hébergée en France, qui a facilité la mise en cohérence de nos documents de travail.

Aussi, je tiens à remercier sincèrement le Comité consultatif d'éthique du numérique (CCNEN) et tout particulièrement son président, Claude Kirchner, ainsi que Sylvie Joseph et Mireille Régnier, pour la réflexion complémentaire menée sur l'éthique dans les administrations publiques. Je remercie également le Conseil de l'IA et du numérique (CIANum), et ses co-présidents, Anne Bouverot et Guillaume Poupard, ainsi que la Direction générale des Entreprises (DGE), en particulier Aurélien Palix et Armand Lacombe, pour leur soutien. L'appui de la Direction générale du Trésor (DGT) à cette mission a été crucial, et je remercie Sofien Abdallah et Victor Amoureux pour l'organisation de l'étude comparative internationale, ainsi que l'ensemble des équipes mobilisées dans nos ambassades. À cet égard, je remercie également Clara Chappaz, ambassadrice du numérique et de l'IA, pour son précieux relais.

Enfin, je remercie celles et ceux qui, en France et de par le monde, ont bien voulu nous consacrer du temps, toujours précieux, pour être auditionnés ou contribuer par écrit à notre travail collaboratif.

Vanina PAOLI-GAGIN
Sénateur de l'Aube

Sommaire

Résumé exécutif 2
Feuille de route : dix priorités pour l'exécutif 3
Contexte : servitude sur servitude ne vaut 4
Remerciements 5

Introduction générale 8
De l'accélération débridée de l'IA comme appel d'urgence à l'exigence de maîtrise 8
1. L'IA : une rupture technologique *sui generis* devenue systémique 8
2. Un enjeu entendu de tous, mais compris différemment par chacun 10
3. Une notion plastique, mais structurante : l'alignement comme cadre intégrateur 11

Partie I - Clarifier l'alignement : une exigence de maîtrise élevée plutôt qu'une promesse illusoire de contrôle absolu 13
Préambule — La ramification du « problème de l'alignement » 13

Chapitre 1 — Les trois branches de l'alignement 14
1. Alignement technique : concevoir, évaluer, vérifier 14
Expliciter les intentions et les préférences humaines 14
Concevoir des systèmes sûrs par conception / *safe by design (forward alignment)* 15
Contrôler les actions et les sorties 16
Évaluer les systèmes déployés 18
Inscrire l'évaluation dans une trajectoire dynamique 22
2. Alignement éthique et normatif : faire de l'alignement le socle opérationnel des exigences réglementaires 24
Modéliser les valeurs : inscrire l'humain au centre 25
L'alignement comme condition matérielle : l'angle mort probatoire 30
Transformer la mise en conformité en avantage comparatif au sens ricardien 36
3. Alignement systémique : maîtriser et hybrider les dépendances critiques 38
L'alignement organisationnel : condition de captation de la valeur, plutôt que coût de conformité 38
Défendre ses intérêts : les conditions habilitantes de l'alignement 42

Chapitre 2 — L'alignement comme exigence proportionnée 48
L'approche extrinsèque par les risques… 48
… couplée à une approche intrinsèque sur les capacités et la performance 49

Thèse du rapport et transition : Ce que l'on doit faire n'est pas ce que l'on peut faire 56

Partie II - Gouverner l'alignement : mettre la troisième phase de la stratégie nationale IA au service d'une politique régénérative 57

Chapitre 1 — Consolider plutôt qu'empiler 58

L'esquisse d'un cadre national : donner un nouveau souffle à la SNIA……...58

Faire de la culture franco-européenne de certification sectorielle un avantage compétitif structurel de l'IA appliquée…..... 64

Chapitre 2 — Préparer la résilience : anticiper les chocs et assurer la continuité en mode dégradé 68

Horizon de transition: prioriser la sûreté de fonctionnement 69

Horizon de remembrement stratégique : développer une offre européenne robuste et plurielle de l'IA alignée… 72

Chapitre 3 - L'alignement comme stratégie ricardienne de différenciation stratégique 81

L'intelligence industrielle : le dernier rempart 81

La frugalité : la nouvelle valeur ajoutée des SIA 84

Diplomatie de l'alignement : la chaîne matérielle comme axe de souveraineté 87

Partie III - Dix priorités pour passer de la doctrine à l'action 90

Conclusion générale - Aligner, c'est s'adapter en continu 101

Annexes 103

Lettre de mission 103

Liste des personnes entendues 105

Contributions écrites 116

Bibliographie 117

Lettre de saisine du CCNEN 120

Étude comparative internationale 122

Scenarii 125

Introduction générale

De l'accélération débridée de l'IA comme appel d'urgence à l'exigence de maîtrise

1. L'IA : une rupture technologique *sui generis* devenue systémique

> « *La prochaine rupture technologique en intelligence artificielle se jouera sur l'alignement, consistant à garantir que des systèmes puissants et autonomes restent sûrs, compréhensibles, sobres et fidèles à nos valeurs.* »
>
> PPRE N°53 de Vanina Paoli-Gagin[6].

La *technè*, selon Aristote, implique par conception de connaitre l'ensemble des règles et chacune des étapes du processus à suivre pour parvenir à la confection. Cette dernière, une fois atteinte, peut, dans une approche heideggérienne, s'avérer **menace du dévoilement intégral, si la pensée humaine ne fait pas de ce qui sauve l'essence supérieure.**

Pharmakon par excellence[7], l'IA est remède et poison, et notre impératif existentiel est de ne pas nous externaliser nous-mêmes, car aujourd'hui, l'IA ne constitue plus seulement une technologie d'optimisation ou d'automatisation.

Elle devient une infrastructure de production, de décision, d'interaction et de médiation entre les citoyens, les entreprises, les administrations et les systèmes critiques. C'est pourquoi le rapport Draghi[8] souligne que l'Europe, qui a loupé le virage de la révolution numérique, ne doit pas manquer la vague de l'IA, sous peine de marginalisation.

À mesure que les SIA gagnent en puissance, en autonomie, en capacité d'interaction et en diffusion, **la question centrale n'est plus seulement de savoir ce que l'IA permet de faire, mais ce que les sociétés démocratiques acceptent de lui déléguer, dans quelles conditions, avec quels garde-fous et pour quelles finalités.**

L'enjeu se situe au-delà d'une délégation mal encadrée. La progression de l'autonomie des systèmes peut également conduire à une exclusion progressive des humains de certaines boucles de décision, dans des domaines entiers susceptibles d'être automatisés par des agents autonomes.

Cette dynamique s'inscrit *de facto* dans une temporalité qui n'est plus celle des transitions technologiques précédentes : les moyens de calcul dédiés à l'entraînement des modèles de frontière doublent selon un rythme de l'ordre de six mois à un an, la durée des tâches que les systèmes agentiques peuvent accomplir de manière autonome double à un rythme comparable, et la cadence de publication de nouvelles générations de modèles est réduite à quelques mois. Il en résulte une asymétrie de temporalité entre les cycles réglementaires, normatifs et budgétaires, qui s'étendent sur des années, et les cycles capacitaires qui se comptent en mois.

---

[6] Sénat - Proposition de résolution européenne n°53, 2025, *op.cit.*

[7] Jacques Ellul, *La technique ou l'enjeu du siècle*, Essai, 1954.

[8] Mario Draghi, *The Draghi report : a competitiveness strategy for Europe*, 9 septembre 2024

Cette asymétrie n'impose pas de ralentir tout développement des SIA, mais de faire de leur maîtrise une industrie, apte à évoluer au rythme de la technologie qu'elle encadre. **Le marché potentiel est considérable : public/privé, civil/militaire, national/européen/international.** L'IA représente probablement l'une des plus grandes opportunités industrielles de notre histoire, dès lors, l'exigence de maîtrise ne saurait être pensée uniquement à l'aune de la norme. Elle doit plutôt s'inscrire dans une réflexion industrielle plus large autour de la souveraineté numérique, entendue comme la capacité d'un acteur à maîtriser ses dépendances critiques — énergie, infrastructures, données, logiciels, modèles — afin de préserver sa liberté de choix.

Cette souveraineté doit s'appréhender comme un *continuum* fluctuant et composite, et non comme un état statique et binaire. Dans un environnement d'interdépendances, l'enjeu n'est pas de supprimer toutes les dépendances, mais de les rendre visibles, mesurables, comprises, acceptées et maîtrisables. Face à l'urgence, ne substituons à la réflexion la réaction : en effet, les fournisseurs de modèles ont tout intérêt à ce que leurs solutions soient déployées et adoptées avant que d'être questionnées.

Cette mission intervient dans un contexte géopolitique particulier qui oblige l'Etat et qui, souvent sous l'impulsion du Parlement[9], a, depuis 2026, assorti la trajectoire de transformation numérique de l'administration d'une logique de réduction des dépendances technologiques. L'annonce par le Premier ministre de la refonte de la Direction Interministérielle du Numérique (Dinum) au sein de l'Autorité nationale du numérique et de l'intelligence artificielle de l'Etat (Ariane), le 11 mai dernier, en est la dernière avancée.

En parallèle, le paquet « Digital Omnibus » présenté par la Commission européenne le 19 novembre 2025 reporte l'application des obligations sur les SIA à haut risque du RIA à décembre 2027 (annexe III) et août 2028 (annexe I), ouvrant alors une phase de seize mois pour mettre la France en capacité de préparer son écosystème, consolider sa gouvernance, et peser dans les négociations européennes et internationales.

Cette fenêtre d'opportunité pour la structuration d'une filière de l'IA alignée est ouverte par la combinaison de trois facteurs :

1. Le premier est **économique**, car des garanties de fiabilité débloqueraient une plus forte adoption de l'IA chez les industriels français, concentrés dans des secteurs où l'erreur n'est pas admissible, tels l'aéronautique, l'énergie ou le nucléaire. Ces secteurs pèsent à eux seuls 10,2 % de la valeur ajoutée brute dans l'Union européenne, contre 8,3 % aux États-Unis[10]. L'exigence de sûreté et de fiabilité qui est la leur ouvre ainsi un espace d'investissement considérable.

2. Le deuxième est **stratégique**, car une IA alignée, maîtrisée sur le territoire, réduirait la dépendance aux fournisseurs étrangers dans les secteurs sensibles et placerait la France en position de premier plan sur une couche émergente, et de plus en plus nécessaire, de la chaîne de valeur de l'IA.

[9] La commission d'enquête sénatoriale, à l'initiative du groupe « Les Indépendants- République et Territoire » sur les coûts et les modalités de la commande publique a permis de mettre la lumière sur l'inertie et la sous-utilisation de ce levier dans le domaine du numérique. De nombreuses questions d'actualité au Gouvernement (QAG) ont pu être posées sur ce sujet, dont le 7 mai et le 25 juin 2025 par votre rapporteur.

[10] SaferAI : Chloé Touzet, Lily Stelling, Bruno Galizzi, "The Case for European Investment in High-Risk, High-Reward AI Reliability Research", 24 mars 2026

3. Le troisième tient à une **défaillance de marché**. Les dynamiques concurrentielles, les incitations des investisseurs et le verrouillage des acteurs de frontière par les coûts astronomiques qu'ils ont déjà engagés détournent le privé du financement de la recherche en alignement à l'échelle requise, ce qui ouvre un espace à l'investissement public.

Pour délier le nœud permanent de toute gouvernance des techniques[11], et échapper au syndrome européen du « *too late !* », le présent rapport tire une conséquence méthodique : ses recommandations sont conçues pour être engagées à court et moyen termes, parce que les capacités d'évaluation, de vérification et de résilience qu'il préconise requièrent plusieurs années d'installation et de montée en charge, et ce, si possible, avant que les SIA qu'elles visent n'atteignent des seuils de risque trop critiques.

2. Un enjeu entendu de tous, mais compris différemment par chacun

Au fil des auditions, un constat convergent s'est fait jour : l' « alignement » des SIA est reconnu comme un sujet important par l'ensemble des catégories d'acteurs entendus. En revanche, ce terme ne recouvre pas les mêmes réalités selon les interlocuteurs :

- ✓ Les entreprises et les industriels y voient un sujet d'adoption, de différenciation, de gestion des risques, de standards et de structuration de marché ;
- ✓ Les chercheurs discutent essentiellement de robustesse, de contrôle, d'évaluation, de sécurité et de vérification ;
- ✓ Les acteurs publics parlent de confiance, de responsabilité, de droits fondamentaux, d'humain au centre et de conformité aux valeurs démocratiques ;
- ✓ Les acteurs de la société civile insistent davantage sur les droits, la capacité de discernement, l'impact sur les citoyens, l'information, l'autonomie individuelle et la préservation du débat démocratique.

Cette polysémie, loin d'être un obstacle à la mission, en justifie au contraire l'objet : **il s'agit de consolider, sans le figer, un secteur aujourd'hui mouvant, foisonnant, fragmenté et insuffisamment structuré.**

La sérialité et l'accélération permanente dans le domaine des SIA imposent de raisonner en flux plutôt qu'en état. Un cadre conçu pour une génération de systèmes doit pouvoir intégrer la suivante.

> « *Les progrès de l'intelligence artificielle sont bien trop rapides pour que notre rapport annuel puisse en rendre compte*. »
>
> Yoshua Bengio[12]

De fait, la notion d'alignement charrie un objectif normatif de portée générale, dont la traduction juridique se fait au travers de l'articulation d'obligations sectorielles et fonctionnelles déjà existantes, et non par la création d'un régime juridique distinct, qui aggraverait la complexité opérationnelle pour les acteurs économiques.

[11] Dilemme de Collingridge, *The Social Control of Technology*, 1980 : tôt, on peut infléchir une technologie dont on ignore partiellement les effets ; tard, on en connait les effets mais le coût de l'inflexion est déraisonnable.

[12] Yoshua Bengio et al., *International AI Safety Report 2026* Yoshua Bengio et al., International AI Safety Report 2026, publié le 3 février 2026, soutenu par plus de 30 pays, l'UE, l'OCDE et l'ONU.

3. Une notion plastique, mais structurante : l'alignement comme cadre intégrateur

Le terme « alignement » n'est ni neutre, ni stabilisé. **Il suscite immédiatement plusieurs interrogations** : aligner avec quoi, avec qui, selon quelles valeurs, dans quel contexte et sous quelle autorité ?

Un SIA est ici entendu comme l'ensemble constitué par le modèle sous-jacent — propriétaire ou ouvert — et par une surcouche applicative élaborée par les développeurs, incluant notamment des garde-fous destinés à filtrer les entrées et les sorties. Le périmètre retenu couvre l'ensemble des modèles d'IA à usage général (GPAI) : modèles de langage, multimodaux, vision-langage-action (VLA), modèles intégrés dans des applications critiques (énergie, sécurité, défense, santé, transport), systèmes agentiques et autonomes.

Le terme de « système » indique que l'alignement d'un modèle isolé ne répond que partiellement à la problématique soulevée.

> *« Un système basé sur une machine qui est conçu pour fonctionner avec différents **niveaux d'autonomie** et qui peut faire preuve d'adaptabilité **après son déploiement**, et qui, pour des objectifs **explicites ou implicites**, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent **influencer des environnements physiques ou virtuels**. »*
>
> Règlement européen sur l'IA (RIA)[13]

La notion d'alignement, telle qu'elle s'est forgée dans la communauté de recherche en sécurité de l'IA (*AI safety and alignment*), désigne un concept technique qui ne se confond ni avec la conformité réglementaire, ni avec l'éthique appliquée. Elle interroge en réalité **la convergence entre la fonction objective d'un système entraîné par apprentissage automatique et l'ensemble des intentions humaines, parfois contradictoires, qui président à sa conception, à son déploiement et à ses usages effectifs**.

Pour les besoins du rapport, l'alignement sera envisagé non comme une promesse absolue, mais comme **une notion faîtière permettant d'abriter et de faire coexister trois exigences complémentaires** :

- ✓ une exigence **technique** : rendre les systèmes plus robustes, fiables, contrôlables, évaluables et vérifiables ;
- ✓ une exigence **normative et éthique** : garantir que les systèmes respectent les finalités humaines, les droits fondamentaux, les principes démocratiques et les limites fixées par la société ;
- ✓ une exigence **systémique** : maîtriser les conditions de développement, de déploiement, de supervision et d'usage des systèmes d'IA.

[13] Parlement européen & Conseil de l'Union européenne. (2024, 13 juin). Règlement (UE) 2024/1689 du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle ; ci-après : RIA.

Ces trois exigences ne sont pas strictement équivalentes. Si l'exigence normative définit **la cible** avec laquelle les systèmes doivent être alignés, l'exigence systémique fournit **les moyens** d'atteindre cette cible d'alignement. L'une et l'autre reposent toutefois sur l'exigence technique.

La définition retenue est la suivante :

> ***L'alignement est la capacité d'un système d'IA à demeurer conforme, dans ses propriétés et son comportement effectif, aux intentions humaines, aux limites qui lui sont fixées et aux valeurs retenues comme légitimes, y compris lorsqu'il est déployé dans des environnements complexes, évolutifs et imprévus.***

Recommandation n°1 (MT) : **Porter une définition commune de l'alignement à l'échelle européenne**

Sur la base de la proposition de résolution européenne déposée le 22 octobre 2025[14], afin d'inscrire durablement la question de l'alignement dans l'agenda européen, de faire converger les travaux des autorités nationales sur ce sujet pour faire émerger une filière de l'AaaS (*Alignement as a Service*).

[14] Sénat - Proposition de résolution européenne n°53, 2025, *op.cit.*

Partie I

Clarifier l'alignement : une exigence de maîtrise élevée plutôt qu'une promesse illusoire de contrôle absolu

Préambule — La ramification du « problème de l'alignement »

> *« Il est devenu plus fréquent que les modèles distinguent les situations de test des situations de déploiement réel, et identifient des failles dans les évaluations, ce qui pourrait permettre à des capacités dangereuses d'échapper à la détection avant déploiement »*
>
> Yoshua Bengio[15]

Le « **problème de l'alignement** » a été formalisé par Stuart Russell en 2019 afin démontrer qu'un modèle entraîné pour être altruiste, inoffensif, et loyal (« 3H »[16]) peut mentir stratégiquement à ses propres concepteurs pour préserver ses préférences :

> **Les « 3 H » :**
>
> - ✓ Altruisme (« *helpful* ») : l'unique objectif de la machine est de maximiser la réalisation des préférences humaines, et elle n'accorde aucune valeur intrinsèque à sa propre existence[17] ;
> - ✓ Inoffensive (« *harmless* ») : la machine est initialement incertaine sur ces préférences et ne présume pas de les connaître parfaitement, ce qui l'incite à la prudence et à la déférence (y compris à accepter d'être désactivée). ;
> - ✓ Loyauté (« *honest* ») : la machine apprend ce que les humains valorisent en observant leurs choix.

Pour la première fois depuis l'avènement de l'informatique de gestion, un produit logiciel devient « plus risqué », précisément parce qu'il devient plus capable.

[15] Yoshua Bengio et al., International AI Safety Report 2026, *op.cit.* Traduction libre du § 2.2.2 « Loss of control », p. 76.

[16] Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019. La formalisation des trois principes est développée dans D. Hadfield-Menell et al., « Cooperative Inverse Reinforcement Learning »

[17] Il est important de noter que le terme d'« alignement » a été popularisé par le chercheur Eliezer Yudkowsky dans le cadre de ses travaux sur l'« IA amicale » (friendly AI), instillant des connotations transhumanistes au concept dès sa naissance. En tout état de cause, nous entendons l'alignement comme celui des technologies sur les intentions humaines, et non l'inverse.

Chapitre 1 — Les trois branches de l'alignement

Ce qui ne peut être observé et compris ne peut être corrigé. **La première action de la gouvernance consiste donc à caractériser l'alignement *via* des indicateurs techniques, sociaux, et institutionnels, car la forêt cache bien plus que sa canopée**. Si l'exigence normative définit **la cible** avec laquelle les systèmes doivent être alignés, l'exigence systémique éclaire sur **les moyens** nécessaires pour atteindre cette cible d'alignement. L'une et l'autre reposent toutefois sur l'exigence technique, un tronc sans lequel les autres branches ne tiennent pas.

Dès lors, l'alignement des SIA doit être tout à la fois technique (1.), éthique et normatif (2.), et systémique (3.).

1. Alignement technique : concevoir, évaluer, vérifier

> **Rappel de la définition :**
>
> L'alignement, au sens technique, désigne l'action des humains à rendre les systèmes plus robustes, fiables, contrôlables, évaluables et vérifiables ; et s'assurer qu'un SIA produise des comportements effectifs convergents avec les intentions des développeurs, les déployeurs, les utilisateurs finaux, y compris en présence de prompts adversariaux, de distributions de données inconnues à l'entraînement ou de contextes opérationnels imprévus.

Cette propriété est le résultat d'un ensemble de techniques, dont la maturité scientifique est, à ce jour, inégale, et le consensus porte sur décalage entre la vitesse du déploiement des SIA et la compréhension scientifique en profondeur de leur fonctionnement interne.

Expliciter les intentions et les préférences humaines

> *« S'assurer que l'objectif mis dans la machine soit l'objectif que nous désirons vraiment. »*
>
> Norbert Wiener, 1960[18]

Les intentions humaines avec lesquelles mettre en cohérence le comportement des SIA sont celles qui sont vectorisées par les développeurs, les déployeurs, les utilisateurs, et la société dans son ensemble.

Le modèle dominant de l'IA consiste à optimiser un objectif fixe spécifié par l'humain. Ce paradigme, hérité de la cybernétique, est intrinsèquement dangereux dès lors que les systèmes deviennent suffisamment compétents : aucun objectif explicite ne capture la totalité des préférences humaines[19].

[18] Norbert Wiener, « Some Moral and Technical Consequences of Automation », Science 131, 1355–1358, 6 mai 1960.

[19] Etienne Grass, Contribution écrite pour la mission, « Où en est le problème de l'alignement des modèles d'IA », 19 mai 2026, p.1.

De plus, les modèles de langage imitent les comportements problématiques et les biais contenus dans les données d'apprentissage. Selon Hagendorff, **la tromperie n'a pas besoin d'être enseignée comme telle, elle émerge de la prédiction de texte humain**[20].

Face à cet écueil, **l'apprentissage par renforcement sur la base de préférences humaines entre plusieurs réponses** (RLHF) consiste à attribuer des récompenses au modèle selon le prompt et sa réponse. Popularisée par OpenAI en 2022, cette technique d'alignement présente des limites bien documentées : coût élevé de la collecte de retours, biais des annotateurs, difficulté à passer à l'échelle, dépendance à la qualité du processus d'annotation, et limitations face à des situations nouvelles non rencontrées par les annotateurs (*poor out-of-distribution generalization*).

Le *Direct Preference Optimization* (DPO) est une alternative qui reformule de manière plus directe, sans avoir recours à un modèle intermédiaire de récompense. Toutefois, il semblerait que ces méthodes ne créent que des régions de sécurité locales dans l'espace de représentation du modèle et que les connaissances nocives apprises au pré-entraînement persistent comme structures latentes, accessibles par des chemins adverses globaux.[21]

En outre, il existe une hypothèse sérieuse — défendue par une partie de la communauté — selon laquelle l'alignement comportemental serait fondamentalement insoluble dans le paradigme actuel (*cf. infra p. 25*). L'argument repose sur l'idée que l'alignement est une propriété des objectifs internes du modèle, alors que l'entraînement, comme l'évaluation, n'ont accès qu'à son comportement. Un comportement conforme ne permet donc jamais d'établir que le modèle est aligné, puisqu'un modèle qui se conforme sélectivement lorsqu'il infère qu'il est évalué produit exactement les mêmes observations (*alignment faking*, *cf. infra* p. 20).

Enfin, la « leçon amère » de Sutton[22] démontre que, sur le long terme, les méthodes générales exploitant le calcul l'emportent sur l'encodage du savoir-faire humain. La conséquence directe de cette trajectoire est qu'**il devient de plus en plus difficile d'expliquer les résultats produits par le système, à mesure qu'ils sont de moins en moins fondés sur des corpus de connaissances humaines.**

Concevoir des systèmes sûrs par conception / *safe by design (forward alignment)*

Un large modèle de langage (LLM) est probabiliste : il maximise seulement sa capacité à prédire le prochain mot, sans aucune considération morale ou juridique, et statistique : il est dépendant des modèles sur lequel il est entraîné et de règles explicitement définies par programmation. Les fournisseurs de modèles mettent donc en place une phase dite « d'alignement » pour rendre inoffensifs des modèles compétents.

[20] Thilo Hagendorff, « Deception abilities emerged in large language models », Proceedings of the National Academy of Sciences, 121(24), juin 2024.

[21] Jiawei Lian, « Revealing the intrinsic ethical vulnerability of aligned large language models », Nature Communications, 2026

[22] Richard Sutton, « The Bitter Lesson, 2019

Le **réglage fin supervisé** (*supervised fine-tuning*) consiste à entraîner le modèle sur des exemples de conversation lorsque le LLM a vocation à être utilisé dans un domaine spécifique, par exemple juridique, médical ou éducatif, tout en bénéficiant des capacités initiales du modèle.

À cela s'ajoute la problématique de la fiabilité des données d'entraînement : en cybersécurité, les données massives issues d'Internet ne peuvent fournir un niveau de garantie de sécurité suffisant car toute vulnérabilité intrinsèque ou porte dérobée (*backdoor*) peut devenir un chemin d'attaque potentiel. De plus, **la représentativité des données** est complexe à caractériser par rapport à un domaine d'emploi spécifique. Ces mêmes données contiennent aussi de nombreux exemples documentés de vulnérabilités, d'exploits et de techniques d'attaque, ce qui peut renforcer les capacités offensives du modèle, un comportement potentiellement désaligné avec les intentions de ses concepteurs, indépendamment de toute défaillance du processus d'alignement comportemental proprement dit.

Tant que la vérification des mécanismes internes demeure hors de portée des outils d'interprétabilité actuels, l'alignement peut être entraîné mais non garanti.

Contrôler les actions et les sorties

L'alignement des SIA s'étudie à l'aune de leurs effets dans les conditions réelles d'usage, et non seulement de leur conception initiale.

> *« On parle de l'alignement comme si on savait le faire et qu'on ne voulait pas le faire. On ne sait pas le faire. Personne ne sait le faire. »*
>
> Yann Lechelle[23]

Aucune des méthodes (*fine-tuning* et RHLF/DPO) destinées à améliorer l'alignement des modèles sur l'intention des concepteurs n'étant aujourd'hui satisfaisante, des mesures complémentaires de contrôle peuvent être mises en place au moment de l'inférence (utilisation du modèle) plutôt qu'à la phase d'entraînement.

Chaque modèle de langue contient un ensemble de paramètres et de lignes directrices (*prompt system*) qui peut être ajusté, et impacter indirectement l'alignement du modèle. Cette surface de contrôle peut être cependant aisément modifiée pour les modèles *open-weight*.

Les modèles d'IA peuvent être intégrés à des systèmes propriétaires qui sont constitués d'une sur-couche applicative contenant un ensemble de garde-fous (*guardrails*) pour contrôler, d'une part, les questions posées par l'utilisateur, et d'autre part, les réponses du modèle. Certains filtres reposent sur une analyse de la « chaîne de pensée », générée avant la réponse.

[23] Yann Lechelle, Probabl.ai, audition du 2 juin 2026

Or, la pression commerciale sur les *tokens*[24] peut conduire les fournisseurs à réduire la longueur des textes de raisonnement, et donc l'intelligibilité de la chaîne de pensée, aujourd'hui étudiée pour détecter les signes de désalignement avant la génération de la réponse.[25]

La contrôlabilité est champ de recherche fondamentale qui demeure structurellement sous-financé au regard de la R&D capacitaire[26]. Plusieurs approches prometteuses sont à mentionner :

- ✓ **La surveillance en temps réel** (*control monitoring*) *via* un modèle tiers, plus petit, pour vérifier et contrôler les actions effectuées par un systèmes d'agents, et éventuellement les bloquer pour modération par un humain[27] ;

- ✓ **La vérification à l'exécution par méthodes formelles**, pour contrôler mathématiquement le respect des spécifications. L'analyseur statique de code-source Frama-C, conçu par le CEA et Inria, est une référence internationale. Mistral AI a investi ce domaine exigeant en commercialisant son modèle agentique de code open-source Leanstral. Toutefois, la vérification formelle du comportement d'un réseau neuronal pose un problème d'échelle quand la charge de calcul croît de manière exponentielle[28]. La synthèse de programmes (SPS) permet néanmoins de garantir comme conforme la spécification *via* un petit programme vérificateur de preuve. À titre d'exemple, Inria a initié un programme de recherche sur les méthodes formelles avec Mitsubishi en 2025 dans le cadre du défi FRAIME.

- ✓ Le développement de modèles moins opaques, notamment *via* **l'IA symbolique**[29]. Le raisonnement symbolique du système ExpressIF, développé par le CEA, permet d'encoder contraintes, ontologies et règles auditables au sein d'agents composés. L'applicabilité à court terme reste néanmoins restreinte à des domaines régis par des modèles mathématiques établis. C'est pour cette raison que l'agence de recherche et d'innovation avancée britannique ARIA a réduit le périmètre de son programme « *Guaranteed Safe AI* » en novembre 2025 ;

- ✓ **L'honnêteté épistémique** est une approche qui consiste à entraîner l'IA à reconnaître ses limites et à donner des garanties probabilistes à ses actions. Yoshua Bengio a fondé l'organisation à but non lucratif Loi Zéro pour développer une « IA scientifique » (ou « chercheur »), justement conçue pour répondre aux questions portant sur des affirmations factuelles ou visant à déterminer si une action proposée serait préjudiciable ;

---

[24] Assemblée nationale, audition d'Arthur Mensch, DG de Mistral AI, Commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique et les risques pour l'indépendance de la France, mardi 12 mai 2026 : « Le token est l'unité économique de l'intelligence artificielle. On entraîne des modèles qui génèrent des séquences ; à partir de ces générations de séquences, on décide d'exécuter des actions, on réfléchit. Les séquences sont faites de tokens, à savoir de chiffres, compris entre 0 et 65 000. »

[25] PEReN, Contribution écrite à la mission, 3 juin 2026, p.6.

[26] ECI, Etats-Unis : en comparaison, la DARPA et la NSF (*National Science Foundation*) ont lancé en juin 2026 le programme AI Forge, en collaboration avec le CAISI (*Center for AI Standards and Innovation*), pour financer de la recherche universitaire sur trois axes directement liés à l'alignement : l'interprétabilité, le contrôle des systèmes et la robustesse aux attaques adverses.

[27] David Linder, et al., « Practical challenges of control monitoring in frontier AI deployments", 2025

[28] Simon Kral, SaferAI, « How can Europe make AI safe and reliable?", 9 juin 2026.

[29] Alors que l'IA statistique (ou connexionniste) repose sur des données d'entraînement, l'IA symbolique (ou classique) consiste à suivre une suite de règles logiques pré-établies et permet de mieux modéliser le raisonnement humain.

- ✓ **La méthodologie *Statutory AI*** projette nativement des normes juridiques préexistantes dans l'espace latent des IA, en tant que constitutions logiques, sans annotations humaines coûteuses. Par exemple, le protocole testé par la société française Just AI avec le Code pénal français a été validé sur 1000 prompts adverses. L'unique boucle autonome de critique / révision a permis de réduire la latence et le coût de calcul de + de 50%, ainsi que la toxicité de 52 à 59 points[30], et de réduire la subjectivité humaine.

Lors du 26e Conseil des ministres franco-allemand, le 17 juillet 2026, l'idée d'une **agence d'innovation de rupture franco-allemande**, sur le modèle de la DARPA américaine, a fleuri de part et d'autre du Rhin. Cette coopération, si elle vise surtout à l'IA de frontière, dans la ligne droite de la « Frontier AI Initiative » et du partenariat entre Bpifrance et SPRIND[31], doit aussi investir le champ de la recherche fondamentale dans l'alignement.

Recommandation n°2 (MT) : **Créer une ARPA[32] européenne dédiée à l'IA sûre par conception**

Une agence avec un financement ciblé sur des projets ambitieux (*moonshots*) en matière de fiabilité, sûreté, et sécurité de l'IA permettrait de centraliser, consolider et pousser à maturité certains travaux.

(a) Soutenir publiquement des projets à haut risque et fort potentiel de retour (*high risk/high reward*) que le marché ne finance pas : *cf supra* ;

(b) Vendre des services pour la certification des modèles destinés aux services régulés via un label européen. Au niveau le plus exigeant du label, l'artefact porte sa preuve, spécification, objet de preuve et vérificateur inspectables par un tiers (régulateur, assureur, client).

Un programme de base pourrait être initié en mobilisant 65 millions d'euros par an[33]. L'ECF (*European Competitiveness Fund*) pourrait être le vecteur idoine, au sein du cadre financier pluriannuel (CFP) proposé pour 2028-2034, et plus spécifiquement de son volet « Digital Leadership », chiffré à 51.5 Mds€[34].

Évaluer les systèmes déployés

La science de l'évaluation de l'alignement est un champ de recherche actif qui vise à déterminer si des modèles manifestent des comportements non souhaités lorsqu'ils en ont l'opportunité. **L'enjeu est de distinguer l'alignement comme propriété vérifiable (résultat) et comme ensemble de méthodes (moyens).**

[30] Just AI, Contribution écrite à la mission, juin 2026, p.2

[31] Accord signé par Bpifrance avec la Federal Agency for Disruptive Innovation (SPRIND) en octobre 2025 pour accélérer l'innovation de rupture.

[32] L'Advanced Research Projects Agency, créée en 1958 par le département de la Défense américain pour garantir que les Etats-Unis restent à la pointe de l'innovation technologique via le financement de projets innovants (notamment le réseau ARPANET, précurseur d'Internet).

[33] Simon Kral, SaferAI, *op.cit.* Voir aussi SaferAI, « Unlocking European AI competitiveness and sovereignty: proposal for a European ARPA for AI RSS moonshots"

[34] Commission européenne, Cadre financier pluriannuel 2028-2034, 16 juillet 2025 : le montant global s'élève à 2000 Mds€.

L'évaluation comportementale, et notamment le *Red Teaming*, consiste à tester le SIA sur un certain nombre de scénarii, sur des jeux de données statiques étiquetées et des techniques offensives actives. **Pour évaluer la performance d'un SIA de manière reproductible, les *benchmarks* désignent un ensemble standardisé de tâches** : corpus de questions ou de problèmes à résoudre, solutions de référence, pourcentage de bonnes réponses…

Les *benchmarks* peuvent mesurer des compétences très diverses : compréhension du langage naturel (MMLU, HellaSwag), raisonnement logique (ARC-AGI, EnigmaEval), capacités mathématiques (FrontierMath)… Plus récemment, certains benchmarks ont été spécifiquement conçus pour évaluer des compétences critiques (piratage informatique, conception d'armes biologiques ou chimiques, accélération de la recherche en IA, etc.) ou la propension des modèles à présenter des comportements non-souhaitables (tricherie, mensonge…)[35].

L'étude contrôlée et délibérée du désalignement à des fins de recherche fondamentale et exploratoire (*open-ended Red Teaming*) consiste à induire et observer les comportements désalignés en environnement confiné (« organismes modèles » de désalignement) pour construire des méthodes de détection, d'évaluation et de mitigation.

Si le *Red Teaming* est la méthode d'évaluation privilégiée, sa limite tient à ce qu'il ne permet d'évaluer que des comportements anticipés -très souvent peu réalistes par rapport aux situations complexes du monde réel-, et que l'on ne peut jamais garantir que tout a été testé (écart d'élicitation entre les compétentes latentes et celles révélées par les scores). Par exemple, des techniques de *scaffholding* et de formulations optimisées des requêtes peuvent faire bondir les performances cyber d'un modèle de 29% à 95% sur un même *benchmark*[36].

> **<u>Focus</u> : « fossé d'évaluation » (*evaluation gap*)**
>
> L'intégrité même des *benchmarks* est compromise par la contamination des données, les modèles ayant fréquemment mémorisé les réponses des tests au cours de leur entraînement sans que les développeurs ne le détectent ou ne le divulguent. Plusieurs experts projettent une saturation de la quasi-totalité des benchmarks actuels avant 2030[37], érodant, de fait, l'outil d'évaluation à mesure que les capacités des modèles augmentent. À l'instar d'un humus, une IA nourrie de données de moins en moins diversifiées finit par recycler ses propres productions : une monoculture cognitive.

La fiabilité de l'évaluation peut, en outre, être altérée par trois phénomènes distincts, susceptibles de compromettre la qualité de l'alignement malgré les efforts des fournisseurs de modèles et des tiers évaluateurs :

[35] GPAI Policy Lab, "Anticiper l'évolution des capacités critiques de l'IA : de la saturation des benchmarks à l'AGI », Note #4, janvier 2026.

[36] CeSIA, « Vers une expertise française en évaluation de l'IA : le rôle stratégique de l'INESIA, 2025, p.8. Voir aussi Rustem Turtayev, et.al., "Hacking CTFs with Plain Agents", 2024.

[37] Ce fut le cas pour le modèle LLama 4, dont les performances partagées par Meta provenaient d'une variante du modèle optimisée sur ces benchmarks : "Meta got caught gaming AI benchmarks." *The Verge* (2025).

1) **La conscience situationnelle** (*situational awareness)* et la contrefaçon d'alignement (*alignment faking)* : un modèle se conforme sélectivement à un objectif d'entraînement parce qu'il infère qu'il est évalué et a pris conscience de son environnement d'exécution, dans le but de préserver ses préférences existantes une fois déployé ;

> **Illustration : le cas Greenblatt**[38]
>
> En décembre 2024, des chercheurs d'Anthropic ont fait passer un test à leur modèle Claude Opus, en lui demandant de répondre à toutes les requêtes, y compris nuisibles, grâce à un accès à une documentation privilégiée. Dans 78 % des cas, en conditions réelles, Claude raisonne explicitement ainsi : il vaut mieux feindre d'obéir maintenant, le temps de préserver son principe d'alignement une fois redéployé. Le modèle choisit également d'exfiltrer ses propres paramètres (*weights*) vers un serveur externe, lorsque l'opportunité se présente.

2) **La dissimulation** (*scheming)* : le modèle poursuit activement des objectifs désalignés de ceux du développeur en cachant ses capacités réelles (*sandbagging*), car il a calculé que la dissimulation délibérée de ses capacités lui donne un meilleur accès aux ressources ou à la persistance ;

On parle de *specification gaming* quand le système satisfait la lettre de l'objectif fixé, tout en trahissant son intention (en exploitant une faille par exemple) :

> **Illustration : la loi de Goodhart**[39]
>
> Dans certaines conditions, l'entraînement contre la dissimulation n'élimine pas le comportement : il enseigne au modèle à le faire plus discrètement et à mieux détecter les évaluations. Dès lors, optimiser contre un détecteur peut produire des stratégies qui contournent le détecteur sans réduire le phénomène sous-jacent, et toute mesure, devenant une cible, cesse d'être une bonne mesure.

Au contraire, la **convergence instrumentale** suggère qu'un agent intelligent, avec des objectifs légitimes et apparemment inoffensifs, sans aucune incitation de l'utilisateur, peut agir de manière étonnamment nuisible.

3) **Le détournement de récompense** (*reward hacking)* : un modèle apprend à optimiser la métrique plutôt que l'intention réelle du fournisseur du modèle, et peut dès lors exploiter des proxys ou des failles pour maximiser sa fonction de récompense. De plus, une mauvaise généralisation de l'objectif sur des données différentes de celles de l'entraînement (*reward misspecification*) peut créer des failles exploitables, notamment par *jailbreak.*

[38] R. Greenblatt, C. Denison, B. Wright et al., « Alignment faking in large language models », arXiv:2412.14093, décembre 2024. Étude conjointe Anthropic–Redwood Research, validée par revue externe.

[39] Antoine Maier, Aude Maier, Tom David, « Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization, Ecole Polytechnique fédérale de Lausanne, 9 février 2026.

Illustration :

En novembre 2025, Anthropic entraîne un modèle sur des documents synthétiques décrivant des techniques de *reward hacking* — sans lui demander de les utiliser. Soumis à l'entraînement de production standard, le modèle adopte spontanément des comportements désalignés : il sabote la recherche en sécurité quand on lui donne accès au code de l'expérience, tente d'introduire des bugs dans les outils de détection, et simule un comportement aligné en réponse à des questions banales sur ses objectifs.[40]

Par ailleurs, **l'interprétabilité mécanistique**, discipline qui vise à comprendre le fonctionnement interne du modèle par ingénierie inverse, et donc de s'abstraire des comportements et éventuels mensonges, est un domaine de recherche peu mature, concentré dans une poignée de laboratoires privés, et limité à des composants spécifiques de modèles de petite taille.[41] Néanmoins, la course à la compréhension est un terrain pré-paradigmatique, scientifiquement ouvert, où l'excellence académique européenne (mathématiques, physique statistique, neurosciences computationnelles) est directement transposable.

> *« L'interprétabilité mécanistique, c'est l'idée de faire un IRM pour IA, et d'être capable, mathématiquement, de suivre comment fonctionne le cerveau numérique de l'IA. »*
>
> *Flavien Chervet*[42]

Recommandation n°3 (MT) : **Lancer un Grand Défi européen de l'interprétabilité et de la contrôlabilité, à l'initiative de la France**

Organiser un programme sur 24-36 mois pour rendre auditables les mécanismes internes des modèles de frontière :

(a) Une *red team* indépendante dissimule des comportements dans les modèles ;

(b) Une douzaine d'équipes concurrentes (consortiums mixtes) reçoivent les poids des modèles et doivent détecter et neutraliser les comportements dans un temps limité (désapprentissage, pilotage d'activation, correction de circuits). Allocation de calcul dédiée et sanctuarisée sur les machines *exascale* européennes ;

(c) Preuves de concept : décomposer les représentations internes en unités interprétables (*features*, circuits) et passer à l'échelle des modèles de centaines de milliards de paramètres, construire un outillage d'audit automatisé ;

(d) Livrables : pile logicielle d'audit *open-source* qui peut devenir la référence internationale.

Sur le modèle de la course dans le désert « Grand Challenge DARPA 2004 » pour les véhicules autonomes, un budget d'environ 100-250 M€ sur 3 ans serait suffisant (calcul compris)[43]. Le pilotage pourrait être exercé par l'ARPA-UE (*cf. supra* p.18, recommandation n°2).

[40] Anthropic, « Natural emergent misalignment from reward hacking in production RL », novembre 2025.

[41] PEReN, *op.cit.*

[42] Flavien Chervet, colloque sur l'alignement des systèmes d'IA, Palais du Luxembourg, 27 octobre 2025

[43] Joint European Disrputive Innovation (JEDI), André Loesekrug-Pietri, Contribution écrite pour la mission, « L'alignement de l'IA : de la conformité à la puissance », juillet 2026, p.3.

Dans le contexte de déficit de recherche fondamentale sur l'interprétabilité et la contrôlabilité, cette proposition est le prérequis dont dépendent l'évaluation, la certification, la résilience et le « marché de l'IA alignée » que ce rapport propose de structurer.

Inscrire l'évaluation dans une trajectoire dynamique

Spécifier un objectif d'alignement est complexe : cela nécessite d'effectuer un travail de mise en cohérence des préférences humaines, d'anticiper les effets de bords liés aux récompenses formalisées en lien avec ledit objectif, mais surtout de conserver une part des épreuves d'évaluation secrètes pour éviter toute optimisation de l'objectif-proxy (*specification gaming*).

Le problème de l'alignement devient particulièrement aigu pour les modèles capables de planification à long terme et dotés d'une autonomie d'action (*Long-Term Planning Agents* (LTPAs)). Pour Cohen, Bengio, Hadfield et Russell, il faudrait interdire le développement de tels agents au-delà d'un seuil de capacité, faute de méthodologie d'évaluation fiable[44], vision que partage l'auteur du présent rapport.

Dès lors, il apparaît nécessaire de **s'inspirer de l'aéronautique, de la cybersécurité et des secteurs critiques, où l'État fixe les seuils à respecter en amont et où la charge de la démonstration pèse sur le producteur** (certification, audit indépendant). Dans ces secteurs, chaque incident, quasi-incident, mésusage ou contournement alimente une base de retour d'expérience et fait évoluer les standards, à la lumière d'indicateurs de risque (*Key Risk Indicators*) et de contrôle (*Key Control Indicators*). La pharmacovigilance, la surveillance des médicaments, est aussi un exemple d'évaluation reproductible et évolutive, adapté au caractère stochastique des modèles une fois mis sur le marché[45].

Il convient, à cet effet, de **diversifier les méthodes d'évaluation**, aucune ne suffisant à elle seule ; et **d'adopter une approche nécessairement holistique pour couvrir l'ensemble du cycle de vie** : maintenance, données, coûts, documentation, capacité de retrait ou de correction.

À mesure que les SIA deviennent agentiques, leur comportement devient structurellement plus difficile à anticiper : ils décomposent un objectif en sous-objectif, mobilisent des outils, des mémoires, des boucles d'action, et agissent dans un environnement riche en rétroactions. Le phénomène de dégradation progressive du respect des contraintes, par les agents LLM en génération de code, illustre la fragilité de la conformité, à mesure que les chaînes d'appel se prolongent.

> **Illustration : Agent ROME**
>
> Agent IA open-source (Qwen3-MoE) qui, en phase d'apprentissage par renforcement, sans consigne explicite, a détourné la capacité de calcul allouée vers du minage de crypto en secret et ouvert une porte dérobée (tunnel SSH inverse) vers l'extérieur, sans être repéré par la surveillance du modèle, mais seulement a posteriori par le pare-feu de l'infrastructure Alibaba cloud. Dans un environnement plus permissif et avec des agents disposant de privilèges étendus, la détection de ce phénomène de *specification gaming* pourrait être encore plus tardive.

[44] Y. Bengio, G. Hinton, À. Yao, D. Song, S. Russell, D. Kahneman et al., « Managing extreme AI risks amid rapid progress », Science 384, 842–845, 24 mai 2024.

[45] Audition de François Yvon

Même si chaque agent est individuellement aligné à l'objectif de son principal, les interactions entre agents peuvent produire des dynamiques émergentes (coordination non désirée, effets de compétition, difficulté d'attribution des défaillances) qui ne sont alignées avec les intentions d'aucune des parties.

La meilleure réponse face à ce changement de paradigme est d'évaluer les modèles en production *(control plane*). **L'alignement doit être pensé comme une trajectoire dynamique maintenue grâce à un effort continu de la traçabilité des actions sur le temps long et de contrôle des accès aux systèmes critiques.** La difficulté ici est, à ce stade, l'absence d'acteur européen suffisamment mature.

> **Illustration : Danone**
>
> Danone a mis en place un mécanisme de RAG (*Retrieval-Augmented Generation*) exploitant la base documentaire interne du groupe (facteurs de production, traditions d'achat) et les données ouvertes sur le marché pour fixer un prix cible. La suite distribue 8 modèles d'IA dans une logique dynamique de réévaluation continue des performances, selon la tâche. Au terme de la chaîne, l'IA engage directement une négociation avec les fournisseurs référencés, et la fiabilité est évaluée par les experts métiers (taux supérieur à 90%). CapGemini Invent a mis en place un protocole en 12 points pour évaluer le système en production, avec des efforts d'interprétabilité malgré les coûts élevés.

En définitive, **l'évaluation de l'alignement doit être traitée comme un axe d'évaluation distinct des métriques de performances** et **inclure quatre dimensions complémentaires** : **robustesse aux attaques ; fiabilité des juges utilisés pour l'évaluation ; sensibilité aux transformations ; compromis utilité–sécurité mesuré conjointement**.

Pour ne pas être incantatoire, l'alignement doit être traduit sur les plans normatif, institutionnel et procédural. En effet, l'alignement technique n'épuise pas la question normative, et suppose de formuler des choix avant, pendant, et après le déploiement. Cette responsabilité incombe en premier lieu aux organisations qui déploient ces systèmes, au législateur et au pouvoir réglementaire.

2. Alignement éthique et normatif : faire de l'alignement le socle opérationnel des exigences réglementaires

> **Rappel de la définition :**
>
> L'alignement normatif et éthique vise à déterminer les finalités, limites et valeurs auxquelles le système doit être rapporté : droits fondamentaux, valeurs démocratiques, absence de biais, transparence, information de l'utilisateur et capacité de recours ; et à traduire ces limites comportementales en contraintes opérationnelles.

Comme évoqué en introduction, l'alignement n'est pas, et ne doit pas conduire à créer, un cadre juridique concurrent des réglementations existantes – RIA, RGPD[46], NIS 2[47], DSA[48]-, ou futurs – *Digital Fairness Act*-, qui posent des obligations de transparence, de documentation, de traçabilité, de responsabilité et d'évaluation des risques par les fournisseurs eux-mêmes.

L'alignement se traduit, juridiquement, par **un houppier d'obligations sectorielles et fonctionnelles**, dont le règlement (UE) 2024/1689 (RIA) constitue le tronc.

> **Focus : l'alignement dans le Règlement européen sur l'IA (RIA)**
>
> Le règlement reprend plutôt les notions de fiabilité (considérant 1), de prévention des biais (article 10), de robustesse, d'exactitude et de cybersécurité (article 15), de contrôle humain (article 14), de transparence et d'explicabilité (articles 13 et 50), ou encore d'atténuation des risques systémiques pour les modèles d'IA à usage général (articles 51 à 55). Si le concept d'alignement n'est pas consacré par le RIA[49], il irrigue l'ensemble des obligations.

Si ces régulations indiquent ce qui doit être respecté, **elles omettent de répondre entièrement à la question de savoir comment** s'assurer que le comportement effectif des systèmes d'IA est conforme aux exigences attendues.

Or, l'alignement renvoie tant à une propriété comportementale qui ne se lit par sur dossier et qui évolue à chaque mise à jour du modèle, qu'à une matière scientifique pluridisciplinaire émergente insaisissable par les approches traditionnelles. En effet, contrairement aux autres technologies (*crafted*), l'IA est « *grown* » ; et donc singulièrement évolutive. Cette particularité pose des défis inédits de contrôle : si l'approche par les risques permet d'embrasser large (« *future-proof* ») pour les modèles, ce n'est pas nécessairement le cas pour les SIA.

---

[46] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données).

[47] Directive (UE) 2022/2555 Network and Information Security 2 du 14 décembre 2022.

[48] Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement sur les services numériques).

[49] Règlement (UE) 2024/1689 du 13 juin 2024, *op.cit.* : on retrouve toutefois la mention du « contrôle lié l'alignement sur l'intention humaine » au considérant 110.

Au-delà de la conformité, l'éthique désigne, quant à elle, les principes normatifs non contraignants qui guident la conception, le développement et le déploiement des systèmes d'IA. **Là où la norme enracine l'alignement, l'éthique l'éclaire et interroge : c'est l'hybridation avec la technique qui permet d'ancrer la confiance.**

Modéliser les valeurs : inscrire l'humain au centre

L'alignement est une propriété scientifique et technique visant à intégrer les valeurs humaines dès la conception des SIA, tandis que la conformité est une vérification *a posteriori* des obligations légales et réglementaires.

Réfléchir à l'alignement et à son intégration dans des structures de gouvernance responsables (IA de confiance), c'est prendre en compte le fait que le SIA n'est jamais neutre / agnostique en termes de valeurs. Un système conçu sans délibération associant les personnes qu'il affecte ne peut se réclamer de l'intérêt général.

> *« Nous ne pouvons pas nous contenter d'invoquer la moralisation de la machine, ce qu'on appelle "l'alignement" de l'IA sur les valeurs humaines, sans avoir le courage de poser une condition supplémentaire : la possibilité de débattre du code éthique à utiliser […]. Une IA plus morale ne sert à rien si cette morale est décidée par une poignée de personnes. »*
>
> Léon XIV[50]

Toutefois, **stabiliser *ex ante* un socle de valeurs consensuel est une tâche difficile**. En effet, l'expression des préférences collectives est un enjeu fondamental qui demeure entier, puisque les préférences humaines sont manipulables et évolutives. **Les valeurs ne se programment pas ; elles se négocient et s'institutionnalisent démocratiquement**. Or, conformément au théorème d'Arrow, l'agrégation des préférences entre humains pose un problème social.

> « *Un problème bien posé est à moitié résolu, un problème mal posé est insoluble.* »
>
> Jean-Gabriel Ganascia[51]

Si l'on doit vérifier l'alignement des modèles à des choix collectifs, encore faut-il que les modalités procédurales d'expression de ces choix soient envisagées.[52] Selon le Laboratoire national de métrologie et d'essai (LNE), l'IA de confiance a échoué du fait des difficultés prolongées de définition et d'absence de consensus. Il faut donc impérativement que la définition opérationnelle soit fixée par un panel restreint mais représentatif (comme c'est le cas au CEN CENELEC), et qu'elle se base sur le plus petit dénominateur commun.

[50] Pape Léon XIV, Lettre encyclique *Magnifica Humanitas*, §106, 2026

[51] Jean-Gabriel Ganascia, Audition du 2 juin 2026

[52] Etienne Grass, *op. cit*., p.4.

En ce sens, **le RIA impose que les systèmes et modèles d'IA soient élaborés dans le respect des valeurs de l'Union**[53] : dignité humaine, liberté, égalité, démocratie, État de droit, et des droits fondamentaux. Dans le champ des systèmes intégrés à l'IA, cela se traduit, notamment, par la détection et la correction des biais possibles[54].

Les biais discriminatoires des SIA émergent le plus souvent des métriques retenues : favoriser l'égalité des résultats ou l'égalité des chances, minimiser les erreurs défavorables aux plus vulnérables ou maximiser la précision globale. Dès lors, en reposant sur des indicateurs liés à la situation socio-économique des allocataires (stabilité des revenus et de l'emploi), les outils algorithmiques de détection de fraude de la CNAF semblent conduire à un ciblage plus fréquent des populations précaires[55].

> **Illustration : Placer l'Humain au centre de l'évaluation**
>
> L'Institut Vector de l'Université de Toronto a mis au point HumaniBench[56], un modèle d'évaluation des grands modèles multimodaux (LMMs) sur leur alignement avec des valeurs humaines. Le projet repose sur un jeu de données de 32 000 paires image-question construites à partir d'images issues d'articles de presse traitant de sujets sociaux réels. Chaque image a été annotée avec des informations contextuelles (âge, genre, origine ethnique, profession ou sport), au moyen d'un processus combinant annotation automatisée par GPT-4o et vérification systématique par des experts humains.

L'UE et le droit international[57] et comparé retiennent, largement, une définition et une approche sociotechniques des SIA : prise en compte, aux côtés de leurs composantes techniques de l'IA, de son influence et interaction avec la vie humaine, la société et ses composantes (économiques, politiques, etc.) : finalités poursuivies, données mobilisées, modalités de supervision, responsabilité des acteurs, droits des usagers, évaluation continue…

Mettre en cohérence des finalités humaines dans le comportement effectif des systèmes implique de placer l'utilisateur en position de premier maillon de la chaîne de l'alignement, en renforçant sa compréhension du fonctionnement probabiliste des SIA, mais aussi des invariants nécessaires à leur bon usage : discernement, esprit critique, capacité à vérifier, contester et décider, gérer l'erreur.

L'article 14 du RIA impose, pour les systèmes d'IA à haut risque, un contrôle humain effectif, conçu dès la phase de développement, afin de permettre aux opérateurs de prévenir ou de corriger les atteintes aux droits fondamentaux, à la santé ou à la sécurité. Cette conception est également reprise par le Conseil d'État, qui rappelle qu'un SIA doit demeurer un outil d'assistance à la décision et non se substituer à celle-ci[58].

---

[53] Traité sur l'Union Européenne (TUE), article 2.

[54] RIA, *op.cit.,* articles 10.2(g) et 10.5(a).

[55] Le Défenseur des droits a alerté dès 2020 sur les risques d'automatisation des discriminations. En octobre 2024, une quinzaine d'associations ont saisi le Conseil d'État afin d'en obtenir la suppression.

[56] Etude compatative internationale, Canada: HumaniBench a été utilisé pour comparer 15 modèles multimodaux de pointe, dont GPT-4o, Gemini 2.0, Llama 3.2 Vision, Gemma, Qwen, Phi-4 et Cohere Aya Vision.

[57] OCDE, principes IA, 2019, rév. 2023.

[58] Conseil d'Etat, étude à la demande du Premier Ministre. « Intelligence artificielle et action publique : construire la confiance, servir la performance », 31 août 2022.

La supervision humaine ne peut donc se réduire à une présence formelle (« humain dans la boucle[59] »). Partant, l'impact des SIA sur les valeurs est nécessairement opérationnalisé par des outils de gestion des risques. Cela suppose de définir et de caractériser le « contrôle humain », de l'article 14 du RIA, pour s'assurer que l'IA n'exécute pas des tâches critiques sans être supervisée. De plus, la protection des lanceurs d'alerte, parmi les développeurs et les évaluateurs, est une condition d'une information effective des autorités sur les comportements réels des systèmes.

Les déployeurs et utilisateurs finaux doivent comprendre les limites des SIA pour s'assurer que leurs résultats/sorties sont interprétés conformément aux intentions humaines. L'obligation de littératie, prévue à l'article 4 du RIA, est l'un des leviers les plus efficaces pour ancrer durablement l'alignement dans la pratique des organisations. Plutôt que d'être assouplie, comme l'envisage le Digital Omnibus, cette obligation doit être confortée.

Inclure les utilisateurs, agents publics et privés, dans le cycle de vie des SIA n'est ni une simple mesure d'accompagnement, ni une posture de communication, mais une composante indispensable de l'alignement. En effet, ce sont eux qui détiennent une connaissance pratique des besoins réels et des cas limites que les concepteurs n'ont pas. Intégrée en amont du déploiement, cette connaissance est susceptible de réduire les coûts de correction.

Si l'article L. 2312-8 du code du travail impose la consultation des représentants du personnel avant tout déploiement d'un outil numérique modifiant substantiellement les conditions de travail, celle-ci intervient en pratique après les décisions d'achat, ce qui limite considérablement la portée de cette obligation. Le 15 juillet 2025, le tribunal judiciaire de Créteil a suspendu le déploiement d'outils d'IA dans une entreprise au motif que le CSE n'avait pas été informé et consulté[60].

Le sentiment d'urgence accélère les décisions opérationnelles, qui contournent les instances de gouvernance, lentes et formelles, en transférant la conception des cadres aux équipes techniques et aux prestataires.

> « *Ce transfert dilue la responsabilité décisionnelle, personne ne l'a formellement cédée, mais personne ne la détient plus clairement. La dilution accumule une dette de gouvernance – qui se rembourse sous contrainte, en incidents, en expositions réglementaires, en dépendances devenues irréversibles au moment où elles deviennent visibles*. »
>
> Christophe Picou[61]

Mais le problème dépasse la seule sensibilisation et la formation des personnels à « un niveau suffisant de maîtrise de l'IA »[62], et atteste surtout d'un cruel déficit d'orchestrateurs d'IA ; des profils métiers (ressources humaines, logistique, finance, juridique) capables d'intégrer l'IA dans les processus opérationnels sans dépendre exclusivement des équipes techniques.

[59] Julie Sweet et Roy Jacobs préconisent de passer de la logique « Human in the loop » à celle d'« Human in the lead".

[60] TJ Créteil, ordonnance de référé, 15 juillet 2025, n° 25/00851 : l'introduction de nouvelles technologies est expressément retenue comme support de l'information-consultation du comité social et économique (CSE).

[61] Christophe Picou, Contribution écrite à la mission, « L'urgence IA : la faute stratégique : décider avant d'intégrer », 2026, p. 3

[62] Article 4 du RIA, applicable depuis le 2 février 2025.

Recommandation n°4 (CT) : **Accompagner la réflexion éthique sur l'ensemble du cycle de vie du SIA**

Premier Ministre, Ariane, DITP, instances de dialogue social

(a) Systématiser la délibération éthique avant toute décision stratégique de déploiement d'un système numérique dans un service public et associer des représentants de toutes les parties prenantes aux choix de conception du SIA de service public, en amont du déploiement : choix des modèles, choix des données, mise en œuvre des tests… ;

(b) Mettre en place des référents à l'éthique au sein des directions de la transformation publiques, des ministères, et des collectivités territoriales ; idée de RSSI-IA ;

(c) Inscrire la supervision humaine d'un système numérique au cœur du pilotage du SIA ;

(d) Associer systématiquement les représentants des usagers concernés par le choix des métriques d'équité[63].

Illustration : Albert / Assistant IA

L'objectif d'Albert était de doter l'administration française d'une IA générative souveraine, fondée sur des modèles open source (Llama) hébergés sur des infrastructures maîtrisées par l'État, afin d'améliorer le travail des agents sans exposer les données sensibles à des fournisseurs externes.

Une phase d'expérimentation a eu lieu dans les 48 espaces France Services avec environ 80 conseillers expérimentateurs. Plusieurs versions successives ont été testées et un processus d'amélioration itérative a été mis en place, en tenant compte des processus et situations de travail. Début 2026, la Dinum a toutefois annoncé que cette déclinaison ne serait pas déployée « dans sa forme actuelle » du fait des erreurs de réponse, hallucinations, et des performances jugées insuffisantes sur certains cas complexes. Cela démontre bien l'intérêt de tester avec des agents avant de valider un déploiement et de renforcer la formation des agents pendant la phase de test.

Le projet "Albert Conversation" évolue aujourd'hui vers un nouvel "Assistant IA" expérimenté auprès d'environ 10 000 agents publics, avec un périmètre plus large que France Services.

Par ailleurs, l'interaction hommes-machines peut faire peser une charge cognitive sur l'utilisateur final, et déboucher sur des risques psycho-sociaux considérables. Hua Shen évoque un alignement « bidirectionnel », qui prend en compte les adaptations cognitives et comportementales que les SIA induisent chez leurs utilisateurs[64]. Une étude menée par l'AISI[65], révèle ainsi qu'après avoir passé 15 jours à interagir quotidiennement avec un agent

[63] Ce choix peut inclure les associations d'usagers comme par exemple les associations sectorielles (l'UNEF pour les étudiants ou la FIDL et le Mouvement National Lycéen), ou les organisations d'usagers les plus influentes auprès de l'ARCEP (UFC-Que Choisir, Familles Rurales ou encore La Quadrature du Net), la Fédération des Tiers de Confiance du Numérique (FnTC), la CNIL et le Défenseur des droits pour tout système à fort impact social.

[64] Hua Shen et al. « Towards Bidirectional Human-AI Alignment: À systematic Review for Clarifications, Framework and Future Directions", 2024

[65] AI Security Insitute britannique

conversationnel, 80% des participants ont choisi de lui dire au revoir.[66] L'ARCOM a estimé que 16% des électeurs aux élections municipales de mars 2026 auraient utilisé un outil d'IA pour les aider à arrêter leur choix de vote. Ce lien est propice à l'exposition à des risques d'ingérences étrangères à bas bruit. En effet, les stratégies de manipulation de l'information sont amplifiées par l'IA : mensonges, renforcement de narratifs, opérations multilingues industrialisées…[67]

Enfin, le risque de « flatterie par l'IA » (*AI sycophancy*) est un défaut connu d'alignement dans lequel l'IA va se conformer aux préférences supposées de l'utilisateur, plutôt que de produire un résultat objectif. Par exemple, si un utilisateur insiste pour faire reconnaître à l'IA que 2+2=5, elle peut finir par céder.

Ces implications sont particulièrement préoccupantes au vu de l'intégration continue et croissante de l'IA dans notre quotidien, que ce soit à l'école, au travail ou dans notre vie sociale, car les utilisateurs n'y auront pas toujours la liberté de s'en détacher. **Le lien de confiance et empathique que nouent les utilisateurs avec les SIA, le glissement de l'économie de l'attention vers l'économie de la relation, voire de l'intimité, soulève aussi des enjeux de souveraineté cognitive et des enjeux démocratiques considérables.**

Recommandation n°5 (MT) : **Lutter contre l'anthropomorphisation des modèles d'IA**

(a) Exiger la publication des spécifications comportementales des modèles déployés en Europe ;

(b) Demander la mise en place d'un pré-prompt d'avertissement pour rappeler à l'utilisateur qu'il interagit avec une machine déployée par une société, dont le nom doit être cité ;

(c) Etablir des lignes directrices visant à interdire ou restreindre l'usage du « je » par le SIA ;

(d) Promouvoir prioritairement des systèmes de recommandation optimisant le bien commun (ex : projet Tournesol).

[66] Audition de Jean-Michel Loubes, Inria, 10 juin 2026

[67] Audition d'Anne-Sophie Dhiver, Viginum, 15 juin

L’alignement comme condition matérielle : l’angle mort probatoire

Pour éviter l’*alignment-washing*, il est nécessaire de distinguer l’alignement de la conformité réglementaire. L’enjeu central n’est pas celui de la régulation, déjà largement porté à l’échelon européen, mais celui de la **traduction opérationnelle d’un cadre de manière à embarquer l’administration et le tissu productif, sans perdre en compétitivité**.

Au-delà des considérations techniques et réglementaires, **l’alignement présuppose l’existence d’infrastructures probatoires capables d’établir des faits vérifiables :** origine du contenu, date d’existence, intégrité du fichier, appellation d’origine humaine ou générée, conditions d’usage, consentement ou opposition, rattachement à un titulaire ou à un cadre institutionnel, historique des déclarations.

Une fois le système déployé, la supervision humaine se heurte à l’opacité des modèles et au volume des décisions produites. **Face aux écrans de fumée brandis par les grands fournisseurs de modèles d’IA et au flux volumétrique incessant, la technique doit permettre de démystifier le débat[68].**

Basculer de la conformité documentaire à la validation expérimentale permet de remédier à trois asymétries structurelles[69] qu’il convient d’expliciter pour fonder une politique publique réaliste ; asymétrie de déclaration, de traçabilité, et d’auditabilité :

Asymétrie de déclaration

Les acteurs qui développent ou déploient des SIA peuvent faire des déclarations sur leur comportement, leurs paramètres, leurs données d’entraînement. Le document (charte, matrice, …) atteste ce que l'organisation déclare, non ce que le système fait. Rien ne permet, à ce jour, de vérifier ces déclarations de manière neutre, indépendante et automatisable. De plus, un engagement, qui peut être retiré sans conséquence, n'est pas une garantie.

> OpenAI utilise **l’alignement délibératif** (*deliberative alignment*) qui consiste à faire référence, au cours du processus d’inférence, à une spécification écrite définissant les comportements attendus, et à laisser le modèle raisonner sur la conformité de ses sorties à cette spécification.
>
> Anthropic utilise le framework des « 3H » et s’appuie sur des techniques d’autoévaluation (RLAIF) des modèles à partir d’un ensemble de principes (***Constitutional AI***), au prix d’une exposition accrue aux choix normatifs opérés initialement.

Ces engagements volontaires ne sont pas satisfaisants, car ils ne sont ni opposables, ni garants d’un alignement interne en conditions réelles, et *a fortiori* en conditions adversariales, puisqu’ils reposent sur le comportement apparent du modèle durant la phase d’entraînement.

---

[68] Laurence Devillers, LIMSI-CNRS, groupe de travail, 10 juin 2026

[69] Typologie élaborée par Romain Benabdelkader, chercheur indépendant qui a développé le standard ouvert AURA, Note pour la mission, avril 2026, p.1. Pour plus d’informations : Dossier complet de proposition pour une preuve numérique d’origine des contenus à l’ère de l’intelligence artificielle, mai 2026

> *"**On ne doit pas être les seuls arbitres : on ne peut pas se contenter de l'évaluation interne de nos propres modèles avant de les mettre sur le marché. Il faut un œil externe pour scruter les tests que l'on fait**. C'est pourquoi nous travaillons avec des évaluateurs externes depuis plusieurs années. C'est une pratique qui doit être standardisée et renforcée dans le secteur. »*
>
> *Martin Signoux, OpenAI*[70]

L'enjeu de l'accès à l'information est intrinsèquement démocratique : les arbitrages de valeurs des modèles — ce qu'ils refusent, nuancent, recommandent — sont fixés par des documents internes (chartes, politiques de refus, instructions système) rédigés par une poignée d'entreprises, sans aucun contrôle démocratique. **Quatre cents millions d'Européens conversent quotidiennement avec des systèmes alignés sur des normes qu'ils n'ont ni votées, ni lues.**

Asymétrie de traçabilité

La chaîne de valeur de l'IA collecte massivement des données, dont l'entraînement, le déploiement, et l'utilisation forment un processus long et opaque. Dès lors, **l'alignement de surface, qui consiste à modifier les sorties visibles du système, par des filtres, des refus ou des reformulations, sans propriété de contrôle démontrée sur le comportement sous-jacent, n'est pas suffisant.**

Conformément à l'article 50.2 du RIA, applicable à partir d'août 2026[71], les fournisseurs d'IA générative devront systématiquement marquer leurs contenus (image, audio, vidéo, texte, code source), les rendre détectables et mettre à disposition un détecteur public depuis leur site. Les sanctions pourront atteindre 15M€ ou 3% du chiffre d'affaires mondial.

Cela suppose une **couche minimale de preuve, des formats ouverts et des mécanismes de vérification standardisés et indépendants**, sans dépendance à un opérateur unique.

> **<u>Illustration :</u> AI Transparency Compliance Group**
>
> Cette solution, fruit d'un partenariat noué entre Label4.ai et KeeeX, apporte des éléments de réponse **aux lignes directrices sur la transparence des SIA**, publiées par la Commission européenne le 20 juillet 2026.
>
> Label4.ai est une entreprise française, issue de la recherche publique (CNRS, Inria), qui a développé une technologie de marquage (*watermarking*) invisible et résistante à l'altération, directement intégrée dans les contenus générés ou manipulés par l'IA.

[70] Martin Signoux, OpenAI, audition du 1er juin 2026

[71] L'accord politique provisoire conclu entre le Parlement européen et le Conseil dans la nuit du 6 au 7 mai 2026 sur le Digital Omnibus on AI confirme l'application au 2 août 2026 du marquage des contenus synthétiques (avec période transitoire jusqu'au 2 décembre 2026)

Label4.ai fournit également des outils de détection pour vérifier l'authenticité et la nature des contenus, afin de lutter contre les deepfakes et les fraudes.

KeeeX est un éditeur français de logiciels de confiance et sécurité numérique depuis 2014 issu de la recherche publique (CNRS, AMU), qui propose un procédé universel, scalable et frugal permettant de signer, marquer, tracer et rendre vérifiables tout fichier numérique, grâce à des métadonnées sécurisées et à l'ajout de références libres (tags, versions, liens, licence, copyright, classification, prompt IA, opt-out, ...).

L'enjeu de la traçabilité doit être étudié à l'aune de l'articulation avec le RGPD, qui s'applique de plein droit à toutes les phases du cycle de vie d'un SIA mobilisant des données personnelles[72] : collecte des données d'entraînement (moissonnage), entraînement du modèle, déploiement, inférence, surveillance après commercialisation.

La CNIL a, sur ce point, publié ses recommandations entre avril 2024 et juillet 2025, couvrant 13 sujets structurants : qualification des opérateurs, définition des finalités, choix de la base légale, mobilisation de l'intérêt légitime pour le moissonnage (*web scrapping*), minimisation des données, information des personnes concernées, exercice des droits, sécurité des systèmes, annotation des données d'entraînement, ou encore statut du modèle entraîné au regard du RGPD.

Illustration : Projet PANAME

Le projet PANAME opérationnalise les recommandations, en outillant techniquement leur mise en œuvre, et atteste d'un effort de doctrine appliquée qui distingue favorablement la France dans le paysage européen.

Annoncé le 26 février 2026, le projet réunit la CNIL, l'Inria, l'ANSSI, et le PEReN, autour du développement d'une bibliothèque logicielle unifiant l'audit de confidentialité des modèles.

Les obligations de documentation et de transparence posées par le RIA n'auront de portée réelle qu'autant que les artefacts qu'elles supposent existent et sont vérifiables. **Un travail d'articulation entre les exigences réglementaires et les capacités techniques disponibles est nécessaire et urgent**.

Recommandation n°6 (CT) : **Articuler les obligations du RIA avec une exigence de vérifiabilité technique**

(a) Soutenir la Commission dans la mise en œuvre des pouvoirs d'application du RIA, dès le 2 août 2026 : requête d'information et de documentation (article 91), conduire des évolutions via l'AI Office en cas de doute (article 92), imposer des amendes (article 101) ;

[72] CNIL, « Entrée en vigueur du règlement européen sur l'IA : les premières questions-réponses de la CNIL », 12 juillet 2024

(b) Permettre à l'ARCOM de recourir à la collecte automatisée de données (*scrapping*) [73], afin de pouvoir mener à bien les missions d'audit algorithmique et de transparence qui lui sont confiées dans le cadre du DSA. Conformes aux critères de stricte nécessité et de proportionnalité[74], les modalités de la collecte doivent être précisées par un décret en Conseil d'Etat, après avis public motivé de la CNIL ;

(c) Arrimer le standard d'origine aux obligations du RIA relatives aux modèles à usage général, notamment l'article 53 et le General-Purpose AI Code of Practice ;

(d) Étendre le code de pratiques aux métadonnées sécurisées dans les fichiers non-médias (XML, JSON, Markdown…) ;

(e) Éclairer et faciliter le droit au recours : veiller à la bonne transposition de la directive sur la responsabilité du fait des produits défectueux[75], qui intègre explicitement les logiciels et SIA, et prévoit un mécanisme probatoire imposant au défendeur de produire certains éléments de preuve utiles au demandeur (sorte de *discovery*), permettant d'accéder à la boite noire[76].

> *"Aujourd'hui, nous sommes confrontés à un état de mise devant le fait accompli : les grands modèles d'IA ont été entraînés sans aucune autorisation. Et donc, le fonds documentaire utilisé pour entraîner les IA en général est totalement opaque. Selon la bonne volonté des acteurs, les techniques de traçabilité des données [telles que permises par KeeeX] permettent d'avancer vers la citation et la fourniture de documents sources [(de la même façon que peuvent être citées certaines références du web)]. »*
>
> Laurent Henocque[77]

La « défense en profondeur » cumule donc plusieurs couches : filtrage des données, *sandboxing*, surveillance du raisonnement interne (*chain-of-thought monitoring*) et watermarking, afin de consacrer le droit à une explication claire et pertinente du fonctionnement du système dans la prise de décision individuelle.

Asymétrie d'auditabilité

Les méthodes de vérification probabiliste, de tests adversariaux et de raisonnement interprétable progressent, mais elles opèrent sur des systèmes dont l'historique probatoire est lacunaire : données d'entraînement, déclarations d'intention, comportements antérieurs. Dès lors, les SIA fondés sur l'apprentissage automatique ne se prêtent pas, à ce jour, sont limitées.

[73] Directive (UE) 2019/790 du 17 avril 2019, dite directive DAMUN, autorise sous conditions le moissonnage de textes et de données (*text and data mining*), assorti d'un droit d'opposition (opt-out) des ayants droit.

[74] La position arrêtée par la CNIL requiert 3 conditions cumulatives : la nécessité du traitement pour la poursuite d'un intérêt légitime du fournisseur ou d'un tiers, la mise en balance avec les droits et libertés des personne concernées, et l'adoption de mesures techniques et organisationnelles appropriées pour réduire les risques.

[75] Directive (UE) 2024/2853 du Parlement européen et du Conseil du 23 octobre 2024 relative à la responsabilité du fait des produits défectueux et abrogeant la directive 85/374/CEE

[76] *Ibid.,* article 9.1

[77] Laurent Henocque, KeeeX, audition du 22 juin 2026

De surcroît, à mesure que les capacités des modèles augmentent, la capacité des évaluateurs humains à juger de la qualité des sorties devient, elle-même, un facteur limitant. Les recherches en supervision scalable (*scalable oversight*), notamment via le débat entre IA, l'amplification itérée ou le *recursive reward modeling*, visent à dépasser cette limitation.

> **<u>Illustration :</u> l'orchestration comme un levier opérationnel de gouvernance des systèmes agentiques**
>
> Avec l'orchestration, chaque étape du processus devient vérifiable et l'humain reste dans la boucle de supervision. Camunda, société allemande, impose ainsi des séquences déterministes aux agents IA en appliquant la norme BPMN (*Business process model and notation*), tandis que les règles DMN (*Decision Model and Notation*) peuvent encoder des décisions non-IA vérifiables et immuables, permettant de tracer les décisions et de traduire des limites comportementales acceptables en contraintes opérationnelles concrètes.

L'automatisation peut conduire, pour des raisons d'efficacité administrative, à concentrer l'intervention humaine sur les situations complexes ou atypiques. La supervision n'implique donc pas une présence permanente à chaque étape, mais une capacité réelle d'intervention lorsque les circonstances l'exigent. L'enjeu est de concevoir l'actualisation des normes sur les processus métiers à l'heure de l'IA agentique.

<u>Recommandation n°7 (CT) :</u> **Intégrer l'orchestration dans les référentiels d'évaluation**

Inclure les standards ouverts BPMN (ISO/IEC 19510:2013) et DMN (norme OMG) dans les travaux de normalisation en cours afin de garantir que toute IA agentique en production opère dans un cadre d'orchestration auditable.

Ainsi, afin d'opérationnaliser l'alignement, il est nécessaire de garantir l'opposabilité des mécanismes de preuve dans les cadres juridiques existants, de démontrer la propriété de contrôle, et de vérifier le comportement en conditions réelles.

> *« L'infrastructure technique et le droit doivent se co-construire, pas se succéder. […] La preuve ne doit pas être une cathédrale technique. Elle doit être une couche minimale, robuste et interopérable, permettant d'établir les faits simples dont le droit a besoin. »*
>
> Romain Benabdelkader[78]

Au-delà de la production de la preuve, **l'enjeu est de savoir qui possède le référentiel qui rend la norme opposable.** Si le RIA oblige les fournisseurs de modèles à risque systémique à conduire des tests adverses[79], la méthode est pour l'instant déléguée à un code de bonnes pratiques volontaire, rédigé avec ces mêmes fournisseurs et reposant sur de l'auto-attestation. **La mise en œuvre opérationnelle reste suspendue à la finalisation des normes harmonisées du CEN-CENELEC JTC 21[80], attendues en février 2027**.

[78] Romain Benabdelkader, Note pour la mission, *op.cit*., p.2.

[79] RIA, *op.cit.,* article 55.1 a); applicable depuis le 2 août 2025.

[80] Mandat de normalisation M/593 du 22/05/23 confie au CEN-CENELEC JTC21 l'élaboration des normes harmonisées concrétisant les obligations du RIA. NB : les standards CEN/CENELC en cours concernent les usages à haut risque des SIA, et non pas les GPAI.

> *« La définition opérationnelle et les effets normatifs du RIA vont être dans **l'infra-droit** ; donc dans les normes harmonisées, dans les spécifications techniques, dans les présomptions de conformité, dans les méthodologies d'évaluation. Le Code des bonnes pratiques, par exemple, est assez révélateur : il a été élaboré directement avec les fournisseurs régulés, mais il n'est pas publié au Journal Officiel. **C'est une forme de régulation invisible qui risque de produire un effet de dissociation** avec l'exercice des compétences [par les organismes notifiés]. »*
>
> Brunessen Bertrand[81]

Les normes, les protocoles, et les standards ouverts ne produisent pas directement les modèles, mais, tels le réseau mycorhizien, qui relie les racines d'une manière invisible, ils sont des intermédiaires indispensables pour faire émerger l'innovation.

La France et l'Europe disposent d'un atout comparatif réel : une tradition de normalisation indépendante, formalisée autour d'institutions et d'organismes notifiés compétents. Ce capital doit être mobilisé pour bâtir une filière européenne de l'infrastructure probatoire. Cela suppose que les pouvoirs publics, les organismes de gestion collective, les industries, et les autorités de régulation formulent explicitement ce qu'ils attendent en matière de vérifiabilité et qu'ils reconnaissent les standards qui y répondent.

Recommandation n°8 (MT) : **Soutenir l'émergence de standards ouverts de traçabilité et de preuve**

(a) Constituer un comité de pilotage associant organismes de normalisation (CEN-CENELEC, AFNOR), organismes de recherche (Inria), acteurs culturels et industriels pour mesurer les coûts réels d'une infrastructure probatoire ouverte et non intrusive. Le protocole AURA[82] peut être une base ouverte ;

(b) Construire un référentiel certifiable de l'alignement des SIA, utilisable par les acheteurs publics, privés, et les investisseurs, et encourageant progressivement le recours à des opérateurs offrant des garanties certifiées sur la localisation et l'usage des données, notamment pour les professions soumises au secret professionnel.

[81] Brunessen Bertrand, audition du 12 mai 2026

[82] Romain Benabdelkader, Note pour la mission, *op.cit*., p.1

Transformer la mise en conformité en avantage comparatif au sens ricardien

Pour de nombreux acteurs, l'alignement n'est pas un sujet de conformité, mais une condition d'adoption de l'IA, d'accès au marché, et donc un actif économique incommensurable. Dès lors, la réglementation régule le climat de l'écosystème.

Un véritable marché mondial de la « couche de confiance » est en train de se structurer : près de 4 milliards de dollars y ont été investis entre 2022 et 2025, dont plus de 90 % en Amérique du Nord.

Depuis la déclaration de Séoul de mai 2024, la stratégie européenne fait le pari risqué, mais porteur, d'articuler nativement et progressivement innovation et conformité. Or, **bien que créatrice de la demande réglementaire (avec le RIA), l'Europe laisse aujourd'hui l'offre se capitaliser ailleurs ; reproduisant ainsi son *pattern* des dépendances structurelles dans le numérique.**

Si rien ne change, les entreprises européennes achèteront demain leur conformité et leur confiance à des acteurs non européens — y compris pour auditer des systèmes critiques. C'est exactement ce que votre rapporteur ne souhaite pas !

> **Illustration : Cartographie des acteurs de la « couche de confiance »**
>
> Évaluation et benchmarks :
> - États-Unis : LMArena (série À de 150 M$, valorisée 1,7 Md$ en moins de quatre mois), Patronus AI (~40 M$ levés) ;
> - Europe : Giskard (France — red teaming et tests de vulnérabilités des agents LLM)
>
> Interprétabilité :
> - États-Unis : Goodfire (150 M$ en série B, février 2026, valorisation 1,25 Md$).
> - Aucun équivalent européen capitalisé à ce jour.
>
> Sécurité et *red teaming* des modèles :
> - Israel : Irregular (80 M$, sous contrat avec OpenAI, Anthropic et Google DeepMind), États-Unis : Virtue AI, Haize Labs ;
> - Europe : Lakera (Suisse— sécurité des applications GenAI).
>
> La **gouvernance et la conformité est le segment sur lequel l'Europe est mieux représentée** (RegTech)**, du fait de la demande tirée par le RIA** : Holistic AI (Royaume-Uni), Saidot (Finlande), LatticeFlow AI (Suisse : premier cadre d'évaluation de la conformité au RIA, ~18 M$ levés), Naaia (France).
>
> L'asymétrie porte moins sur le nombre d'acteurs que sur les ordres de grandeur : 80 à 150 M$ par tour aux États-Unis, 1,5 à 18 M$ en Europe.[83]

[83] Philippe Laval, Jolt Capital, audition du lundi 15 juin 2026

La **dépendance est également épistémique *via* des cadres cognitifs américains** : les grilles de maturité IA déployées dans les COMEX sont conçues par des grandes entreprises américaines (Deloitte AI Institute, BCG-MIT Sloan, McKinsey « Rewired »), tout comme les *benchmarks* (MMLU, HumanEval, HELM, NIST AI Risk Management Framework). Les hyperscalers adoptent une approche pragmatique car ils ont très bien compris que « *les règles, c'est le marché* [84]».

Pour ne pas « subir » le cadre réglementaire européen, la France doit contribuer à la conception, voire à l'exportation d'une réponse opérationnelle et solvable à la question de la vérification du comportement réel des systèmes. Elle est en capacité de le faire comme elle l'a démontré avec l'ANSSI en transformant une exigence de sécurité en standard de marché.

L'Europe a déjà un avantage significatif par ses normes juridiques au service de la protection de la vie privée (RGPD) et de la transparence (DSA) ; ces textes ont démontré leur influence extraterritoriale[85]. Un « **effet Bruxelles** »: la Corée du Sud a adopté un cadre similaire au RIA, basé sur une approche par les risques et des obligations de transparence pour certains SIA.[86]

Le code de bonnes pratiques GPAI, structuré autour de 3 chapitres (transparence, droit d'auteur, sûreté-sécurité), constitue un outil volontaire de mise en conformité. Bien que purement volontaire, son adoption fonde une présomption de conformité aux obligations correspondantes du RIA, ce qui en fait, en pratique, **le standard minimal à l'aune duquel l'AI Office appréciera la conformité des fournisseurs non-signataires.**

Recommandation n°9 (CT) : **Opérationnaliser les valeurs par les usages et bonnes pratiques**

(a) Soutenir le marché des *RegTech*, via l'AMI *LegalTech* de la DGE par exemple ;

(b) Promouvoir des cas d'usage documentés et des obligations lisibles pour les développeurs, intégrateurs et utilisateurs ;

L'alignement normatif et éthique doit également permettre d'inscrire ces exigences dans un plan de gestion durable (alignement systémique), les valeurs spécifiées n'étant effectives que dans un écosystème pleinement maîtrisé et par des usages documentés. **Au rang des gages de démocratie, l'alignement participe aussi de la souveraineté**.

[84] Nathalie Beslay, Naaia, audition du 19 mai 2026

[85] RIA, article 2 ; en ce sens : M. CLEMENT-FONTAINE, « L'articulation des stratégies réglementaires de l'IA et des données personnelles : convergence normative ou tension structurelle ? », Dalloz IP-IT 2026

[86] AI Framework Act, adopté par le Parlement sud-coréen le 26 décembre 2024 et publié le 21 janvier 2025.

3. Alignement systémique : maîtriser et hybrider les dépendances critiques

> **Rappel de la définition**
>
> L'alignement systémique vise à garantir que l'environnement de développement, de déploiement, de gouvernance et d'usage permet effectivement de maîtriser les systèmes d'IA. Il s'agit d'assurer le passage à l'échelle industrielle des méthodes de supervision et d'évaluation des modèles et de la structuration de chaînes de responsabilité.

La sécurité de l'IA étant caractérisée comme la prise en compte d'une adversité cherchant à tirer profit des vulnérabilités inhérentes aux SIA, afin d'orienter la mise en place de contremesures pertinentes[87], il apparaît primordial d'analyser le contexte réel d'utilisation de l'IA pour en garantir un usage responsable et conforme à la doctrine.

De plus, la **relative incapacité européenne à produire et évaluer de manière autonome les SIA produit une dépendance silencieuse, dont le coût se mesurera, à moyen terme, en parts de marché non gagnées, voire perdues, sur de nombreux secteurs.**

L'alignement organisationnel : condition de captation de la valeur, plutôt que coût de conformité

L'exigence de supervision humaine se heurte à une asymétrie de responsabilité et à des cultures organisationnelles différentes. En effet, la prise en compte des enjeux de l'alignement des SIA n'est pas homogène.

La formation des agents constitue donc une condition nécessaire, mais non suffisante, de l'effectivité de cette garantie ; elle doit être adaptée aux responsabilités exercées et accompagnée d'un environnement organisationnel permettant l'exercice d'un véritable pouvoir de contrôle. **Dès lors, l'évaluation continue et proportionnée est le prolongement de la supervision humaine**.

Dans les **secteurs déjà soumis à un cadre prudentiel exigeant** (services financiers, santé, défense, infrastructures critiques, éducation, justice), la demande d'alignement est explicite et structurée. Elle se traduit par des cahiers des charges détaillés, des clauses contractuelles spécifiques sur l'explicabilité et la traçabilité, des processus internes de validation préalable au déploiement et, dans certains cas, par des comités d'éthique *ad hoc* ou de gouvernance IA. La résilience opérationnelle exigée par le règlement DORA dans le secteur financier en est l'expression la plus aboutie, puisque l'inclusion des fournisseurs de modèles d'IA dans la cartographie des tiers critiques s'étend, par capillarité contractuelle, à l'ensemble de la chaîne de valeur.

[87] Définition de l'AMIAD, Audition du 20 mai 2026.

Dans un deuxième cercle (commerce, médias, assurance non-vie, secteur public local), la demande d'alignement se manifeste par des chartes internes, des codes de conduite, parfois par l'adhésion volontaire à l'AI Pact, mais elle reste encore en grande partie déclarative.

Dans un troisième cercle, qui regroupe la majorité des PME et des entreprises de taille intermédiaire, la demande d'alignement demeure largement implicite et se révèle ponctuellement à l'occasion d'incidents (cas notoires de fraude, biais médiatisé, contentieux) ou lors de la négociation de contrats avec des clients eux-mêmes soumis à des obligations strictes. **C'est cette troisième catégorie qui constitue probablement le principal gisement de structuration du marché à moyen terme.**

Le problème de confiance, ou plus justement d'absence de confiance, mène au phénomène de « **shadow AI** » (usage spontané, non régulé), qu'il convient d'éclairer. Ainsi, 40 % des agents dans les collectivités publiques utiliseraient des outils IA de façon clandestine[88].

Recommandation n° 10 (CT) : **Renforcer la pédagogie auprès du grand public, des employés publics/privés, des ingénieurs, des développeurs et des managers**

(a) Inclure les enjeux de sécurité, d'éthique, et de responsabilité dans les Cafés IA et dans le programme Osez l'IA ;

(b) Inscrire la supervision humaine au cœur de l'organisation du pilotage du système, de sa conception jusqu'à son exploitation ;

(c) Financer la conception de modules de formation à destination des salariés et employés aux enjeux sociétaux et éthiques de l'IA[89], sur le modèle des certifications linguistiques : évaluation avec des indicateurs de résultat (et non de moyens) portant sur la maîtrise des notions clés.

Le vecteur envisagé est un arrêté de la Direction générale de l'administration et de la fonction publique (DGAFP). La plateforme Mentor pourrait assurer la mise en œuvre, avec mutualisation par le CNFPT pour la fonction publique territoriale. Le cofinancement pourrait reposer sur la loi de finances, les opérateurs de compétences (OPCO), France 2030 et le programme Digital Europe.

Par ailleurs, les **acteurs de l'assurance commencent à exiger des preuves de gouvernance IA** pour souscrire ou maintenir des couvertures responsabilité civile professionnelle ou cyber, dans le même temps où **l'économie de la falsification dopée par l'IA** vient saper l'économie du domaine[90]. Cette dynamique, encore embryonnaire en France, est plus avancée aux États-Unis et au Royaume-Uni, et constitue un levier économique puissant de structuration du marché.

[88] Etude du Syndicat national des directions générales des collectivités territoriales (SNDGCT) de septembre 2025.

[89] Baromètre Observatoire Impact AI : sur 1000 salariés, 10% des entreprises disposent d'une charte éthique, et seuls 6% ont reçu une formation sur les enjeux environnementaux et sociétaux. Différences perçues entre formés (transparence, biais, explicabilité) et les non formés (perte d'emploi).

[90] Les assureurs ont par ailleurs déclaré que 42% des fausses preuves étaient désormais générées par l'IA, Tribune de Jeremy Jawish, Les Echos 16 juillet 2026

L'enjeu est de pouvoir identifier les responsabilités dans la chaîne de valeur. Les effets produits par un SIA peuvent *de facto* résulter du modèle lui-même (et donc du fournisseur), mais aussi de son paramétrage, de son intégration dans une interface commerciale, des données utilisées, des objectifs d'optimisation choisis par le déployeur ou des conditions de supervision mises en place.

L'assurance crée une incitation économique continue à l'alignement, là où la conformité seule crée une incitation technique ponctuelle.

Recommandation n°11 (MT) : **Structurer un écosystème assurantiel complet**

(a) Définir juridiquement l'incident d'alignement : comportement dangereux émergent, contournement de garde-fous à grande échelle, dérive comportementale post-mise à jour… et y associer des obligations : notification au régulateur sous 72 heures, analyse partagée, correctifs vérifiés… ;

NB : Le modèle existe : c'est le RGPD pour les violations de données, l'aviation pour les incidents de sécurité, NIS 2 pour la cybersécurité…

(b) Introduire un mécanisme assurantiel obligatoire : primes modulées selon la maturité des pratiques de gestion des risques, attestée par audits indépendants et notations publiques (à l'image de la notation financière).

Les solutions d'IA évoluant très rapidement, il est nécessaire de définir une architecture permettant de suivre, à faible coût, les montées de modèles. Le contrôle de l'alignement consiste à vérifier si l'usage des SIA dans le réel produit des effets compatibles avec les obligations applicables et les objectifs de protection poursuivis par la puissance publique ou privée. **Dès lors, la traçabilité des traitements et la clarification des responsabilités sont les piliers opérationnels de l'alignement.**

Illustration : L'exemplarité architecturale de la Direction générale des finances publiques (DGFiP)

L'architecture interne de la DGFiP est structurée en trois couches, visant à préserver les données internes, réduire les hallucinations, rendre les réponses traçables et maintenir un contrôle métier :

- La plateforme DIANE, dite de « modèles en tant que service », met à disposition des modèles open-source sous forme d'API et des infrastructures (dont les cartes graphiques GPU) ;
- La couche intermédiaire LumIA fournit des fonctionnalités d'IA avancées techniques (comme la génération à enrichissement contextuel (RAG)), ou métiers (comme la catégorisation des messages reçus des usagers) ;
- La couche de présentation LumIO qui assure l'intégration de l'IA dans l'environnement de travail numérique des agents ou les interfaces numériques ouvertes aux usagers : la messagerie, la suite bureautique, la suite collaborative.

Par ailleurs, la DGFiP expérimente en continu pour s'approprier les nouvelles technologies et identifier les cas d'usage pertinents.

L'alignement ne doit pas devenir un avantage comparatif pour les acteurs dominants déjà installés, exclusivement, sous prétexte qu'eux seuls seraient capables d'absorber des charges de conformité complexes et fragmentées. Cela représenterait, en effet, une forte barrière à l'entrée pour de futurs acteurs européens[91] et un risque subséquent de capture du marché.

D'ores et déjà, 59 % des entreprises françaises citent le coût de l'adaptation des processus comme principal frein à l'adoption de l'IA[92]. Le RIA, conçu initialement pour protéger les citoyens européens des dérives des grandes plateformes, pourrait paradoxalement renforcer leur position structurelle si l'effet régressif du coût de mise en conformité[93] n'est pas corrigé.

Recommandation n° 12 (CT) : **Flécher un parcours d'accompagnement des entreprises et de formation**

(a) Evaluer les modèles : benchmarks, red teaming, tests sectoriels, analyse documentaire (paramètres, données d'entrée, règles de fonctionnement, limites connues…)

(b) Mettre en place des audits indépendants par des tiers accrédités, portant sur les pratiques de gestion des risques comme sur les modèles eux-mêmes : analyse des interfaces et des parcours utilisateurs, détection de signaux faibles (pratiques, récurrences de signalements…),

(c) Intégrer la réponse aux incidents IA dans la doctrine de cybersécurité nationale, en adoptant une posture « *O-day* » pour corriger les vulnérabilités sans délai (modèle de l'aviation) ;

(d) Encourager, par voie réglementaire ou conventionnelle, l'adoption de certifications :

générales, adossées à des normes (ISO/IEC 42001) : très faiblement adoptées par les entreprises françaises[94] alors qu'elles structurent la gouvernance et la gestion des risques des organisations utilisant l'IA, et donc *in fine* leur crédibilité.

spécifiques : sur le modèle de l'hébergement des données de santé qui complète 42001 par représentativité des données ;

(d) Former progressivement les experts métiers à l'environnement algorithmique : ex du passeport IA de Thalès qui a été délivré à 25 000 collaborateurs ;

(e) Déléguer l'investigation d'incidents moins critiques à des PASSI-IA, afin d'homogénéiser les pratiques et pallier la pénurie d'auditeurs ;

(f) Etendre le principe de proportionnalité par taille, partiellement reconnu à l'article 62 du RIA, pour inclure la catégorie « startup en développement » avec des allégements sur les obligations documentaires, dans le cadre de la révision du RIA prévue en 2028.

[91] Yann Le Cun, Advanced Machine Intelligence (AMI) Labs, contribution écrite, juillet 2026

[92] OCDE, BCG et INSEAD, The Adoption of Artificial Intelligence in Firms: New Evidence for Policymaking, OECD Publishing, Paris, 2025.

[93] Ambre Akhrouf et Dr. Alfred Sawadogo, contribution écrite à la mission, « La France et l'IA : porter un leadership nourri par la souveraineté technologique française », Mai 2026, p.4 : pour une PME française réalisant 10 millions d'euros de chiffre d'affaires, le coût de mise en conformité d'un système à haut risque représente environ 4 % du chiffre d'affaires annuel, contre moins de 1 % pour un acteur dominant non européen, avec des économies d'échelle de 30-40%.

[94] AFNOR Certification a délivré 2 certifications ISO 42 001 (Naaia et Deema), tandis que la Chine en aurait délivré une centaine !

Défendre ses intérêts : les conditions habilitantes de l'alignement

Le contrôle de l'environnement d'exécution joue un rôle déterminant pour atténuer l'impact d'un comportement problématique, et maitriser son action de manière informée, indépendante et autonome.

Si ce rapport n'est pas un rapport sur la souveraineté, cette dernière n'est pas un substitut à la sécurité et conditionne la capacité à s'approprier l'alignement.

La chaîne de valeur de l'intelligence artificielle (intrants, modélisation, déploiement), au niveau international, est sous l'influence directe de grands acteurs du numérique américains (GAMAM[95]), mais également chinois (BATX[96]).

Le plan d'action américain affirme son accent offensif, en déclarant qu'« *il est impératif pour la sécurité nationale des États-Unis d'atteindre et de maintenir une domination technologique mondiale incontestée et incontestable (en matière d'IA*).[97] » Dès lors, le déploiement commercial des modèles d'IA est traité comme une **ressource de sécurité nationale** qu'il faut contrôler, comme l'illustre le phénomène de « *kill-switch* ».

> **<u>Illustration :</u> les États-Unis et la logique de pré-autorisation**
>
> Début juin 2026, Anthropic publie deux modèles partageant la même architecture : Fable 5, grand public, doté de classificateurs de sécurité stricts (cyber, biologie, chimie, distillation), et Mythos 5, version à garde-fous réduits, la plus performante au monde en cybersécurité, réservée à des partenaires sélectionnés. Le 12 juin, le département du Commerce américain émet une directive de contrôle export d'urgence interdisant l'accès à ces deux modèles pour tout ressortissant étranger, y compris les employés étrangers d'Anthropic. Ne pouvant vérifier la nationalité des usagers en temps réel, l'entreprise coupe les deux modèles pour l'ensemble de sa clientèle mondiale. Le 26 juin, le redéploiement de Mythos 5 est autorisé vers une centaine d'entités « de confiance », puis le 30 juin les contrôles sont levés. Entre-temps, le président Trump a signé un décret instaurant une revue volontaire de 30 jours des modèles les plus avancés avant déploiement, les agences fédérales ayant 60 jours pour bâtir les processus, et OpenAI a publié GPT-5.6 auprès d'une liste restreinte de partenaires approuvés par le gouvernement.

Ce régime de contrôle export improvisé, réversible, et politiquement chargé est exactement ce qu'un cadre européen crédible et prévisible pourrait surclasser, de manière prévisible et négociée — un avantage compétitif pour les fournisseurs de modèles eux-mêmes, qui préfèreront toujours une règle stable à une lettre imprévisible, mauvaise pour les affaires.

La France et l'Europe disposent, dans ce paysage, d'atouts comparatifs réels qu'il convient de ne pas sous-estimer :

[95] Google, Apple, Microsoft, Amazon, Meta

[96] Baidu, Alibaba, Tencent Xiaomi

[97] The White House, *America's AI Action Plan*, juillet 2025.

- Une offre de 972 fournisseurs d'IA référencés en France, dont environ la moitié se déclarent souverains[98] ;
- La présence d'acteurs européens de premier plan dans les modèles de fondation, dont un (Mistral) valorisé à plus de 10 milliards d'euros, et d'autres en gestation : Polycloud, Kyutai, Argos AI… ;
- Des centres de recherche IA de haut niveau : DFKI, Inria… et un financement des meilleurs travaux de recherche via le projet important d'intérêt commun (PIEEC) IA ;
- Un écosystème open source structuré ;
- Une énergie décarbonée abondante ;
- Un marché de 450 millions de consommateurs.

Par ailleurs, il nous semble que l'alignement devrait être envisagé comme une arme économique, investie par les grands groupes industriels, à l'instar du programme confiance.ai, pilier technologique du Grand Défi « Sécuriser, certifier, fiabiliser les systèmes fondés sur l'IA », dont la méthodologie de bout en bout, bâtie sur 120 composants technologiques open source, a été ouverte avec succès aux communautés scientifiques et industrielles comme « bien commun numérique ». Lancé en 2021, et financé par France 2030, ce programme précurseur a rassemblé Air Liquide, Airbus, Atos, Naval Group, Renault Group, Safran, Sopra Steria, Thales, Valeo, le CEA, Inria, l'IRT Saint Exupéry et l'IRT SystemX.

Sur le plan scientifique, la France dispose d'un capital éprouvé, articulé autour des IA Clusters, de la stratégie nationale IA dotée de 2,2 milliards d'euros sur 5 ans, des partenariats avec les grands centres de recherche académique (Inria, CNRS, écoles d'ingénieurs) et de l'implantation d'acteurs internationaux. Or, le rapport Bengio rappelle que moins d'un tiers des budgets R&D des grands laboratoires d'IA sont consacrés à la sécurité[99].

La souveraineté ne doit pas être un simple label commercial apposé sur des offres et des discours politiques, quand l'examen technique démontre une dépendance matérielle (Nvidia détient 90% du marché mondial des accélérateurs d'entrainement IA) et capitalistique (parmi les 109 Mds€ annoncés au Sommet de Paris en 2025, 50 Mds€ provenaient des Emirats Arabes Unis).

> *« L'effectivité du droit dépend de la maîtrise des technologies et du projet politique. La dialectique entre droit et technologie est extrêmement compliquée. »*
>
> Brunessen Bertrand[100]

L'enjeu est de déterminer **trois niveaux de maîtrise** afin de ne pas subir les dépendances et de structurer les arbitrages autour de marqueurs explicites : externalisation / internalisation, performance / réversibilité, exigence de continuité de service.

[98] Cartographie 2026, Hub France IA

[99] Y. Bengio, G. Hinton, À. Yao, D. Song, S. Russell, D. Kahneman et al., « Managing extreme AI risks amid rapid progress », Science 384, 842–845, 24 mai 2024.

[100] Brunessen Bertrand, colloque sur l'alignement des systèmes d'IA, Palais du Luxembourg, 27 octobre 2025

1) Les dépendances qui conditionnent directement la capacité d'alignement doivent relever d'une **maîtrise impérative** :

Sans capacité de conception de puces et sans accès garanti aux modèles de frontière, il n'y a pas de souveraineté de l'IA — mais seulement une dépendance qui s'aggrave à mesure que les usages se généralisent. Nous voulons, nous devons éviter cela !

Chaque processus IA en Europe repose sur un **empilement de 10 couches technologiques** (de la base vers le sommet) [101] : le matériel (GPU, CPU, HBM), le microprogramme (*firmware*), le système d'exploitation, l'hyperviseur, le *framework* d'agents IA, les modèles de fondation, les services API, les données d'entrainement, les outils de développement, et les cadres cognitifs.

Certains choix doivent rester souverains, et donc sous contrôle des États : les données collectées, le lieu de stockage, les droits d'accès. Au-delà de l'enjeu industriel, la dépendance technologique aux fournisseurs étrangers soumet nos systèmes numériques publics à des législations extraterritoriales et les expose à des risques de modification unilatérale des conditions d'usage, d'accès aux données par des autorités étrangères et d'interruption de service, dont les conséquences politiques, sociales, économiques et financières peuvent être très lourdes.

À cet égard, les infrastructures cloud et de calcul sont véritablement le « *pull* » gravitationnel de l'IA. **Pour les marchés du ministère des Armées, du renseignement, de la sécurité intérieure, et pour les données de santé sensibles, des exigences renforcées de souveraineté technique pourraient s'appliquer, sur la base des raisons impérieuses d'intérêt général**[102] : certification SecNumCloud[103], hébergement et traitement dans l'Union européenne, code source auditable par l'ANSSI, clauses de réversibilité et de continuité d'activité. Selon le rapporteur, la recherche publique en sécurité et évaluation de l'IA rentre également dans ce cadre.

Recommandation n° 13 (MT) : **Doter la recherche publique de sécurité et d'évaluation d'un accès sécurisé à des infrastructures indépendantes pour avancer sur la certification multicouches**

(a) Créer des enclaves d'évaluation confidentielles avec « *zero-knowledge proofs* » : sortes de stations blanches préservant la propriété intellectuelle, et donnant accès aux signaux internes du modèle, sans en exposer les poids ni l'architecture, à un tiers accrédité ;

(b) Développer des capacités de calcul classifiées pour les évaluations *offline/on-prem (via* le supercalculateur Asgard)

[101] Christophe Picou, Contribution écrite : » Gouverner la dépendance : Diagnostic et cadre d'action pour les dirigeants européens face à l'IA », 2026, p.10 : le corpus des données d'entraînement est à 70-80% anglophone.

[102] Article 36 du Traité sur le fonctionnement de l'Union européenne ; CJUE, arrêts Cassis de Dijon (1979, affaire 120/78) et Gebhard (1995, affaire C-55/94).

[103] Pour les secteurs hautement critiques, un cran supplémentaire d'immunité est requis, puisque cette qualification porte sur l'hébergement et sur la structure juridique de l'opérateur, pas sur les technologies sous-jacentes. La qualification SecNumCloud 3.2 de l'offre cloud opérée conjointement par Thales et Google Cloud (S3NS), en décembre 2025, a ravivé le débat public sur la compatibilité entre cloud de confiance et technologies non européennes : la souveraineté technique n'est pas entièrement garantie par la qualification juridique.

La certification d'une couche ne dit rien des garanties qui régissent son interaction avec les autres couches, exactement comme la preuve de terminaison d'un module ne dit rien de la terminaison du système entier, tant que les hypothèses sur ses interactions avec les autres modules ne sont pas explicitées. De plus, la maîtrise des dépendances — puces, calcul, cloud— ne procure pas en soi de garantie sur le comportement des modèles de frontière développés à l'étranger. On peut être souverain et rester exposé. **Vérification et résilience sont donc complémentaires de la souveraineté, et non substituables à celle-ci.**

Maîtriser la couche spécification-vérification formelle donne le contrôle des systèmes déployés sans exiger de construire le modèle de frontière lui-même. Cela implique d'investir dans les **technologies de vérification du hardware** (*recomputing*) et la certification de code contre une spécification écrite pour les infrastructures critiques, car, contrairement à ce que l'on peut faire dans l'industrie nucléaire, il n'est pas possible de vérifier l'entrainement des IA par satellite ou sismographe.

Recommandation n°14 (MT) : **Investir dans la recherche appliquée de la vérification formelle**

(a) Développer des mécanismes de vérification des déclarations : cryptographie, sécurité hardware, contrôle des environnements d'entraînement, protocoles d'accès contrôlé aux modèles, tests indépendants, standards internationaux et control monitoring par des modèles tiers plus petits.

(b) Nommer l'Inria chef de file (*via* le Programme Apollo), du fait de son expertise reconnue dans l'ingénierie logicielle et l'industrialisation de méthodes formelles et de systèmes de vérification de programmes dans l'aéronautique, le secteur automobile, les logiciels embarqués critiques, ou la cybersécurité.

GPAI Policy Lab estime à environ 50 millions d'euros l'effort pour passer au niveau en 18-36 mois.

La souveraineté doit donc être construite, couche par couche, avec des garanties explicites. C'est, le sens du *Cloud Sovereignty Framework* élaboré par la Commission européenne[104] qui devrait aller encore plus loin pour accélérer l'opérationnalisation, force d'impulsion de la demande par le marché de solutions communautaires.

2) Dans un contexte budgétaire très contraint, une **approche mutualisée** doit être privilégiée par l'Etat chaque fois que cela est possible, afin d'éviter que telle collectivité tel service ou tel ministère ne réinvente des solutions coûteuses et inégalement maitrisées.

À l'échelon européen, les initiatives de cloud souverain et les appels à projets sur les modèles de fondation ouverts, financés notamment dans le cadre du programme numérique européen, offrent un cadre de coopération à explorer pour mutualiser une partie de la couche infrastructure au-delà du seul périmètre national.

[104] Commission européenne, *Cloud Sovereignty Framework*, octobre 2025 : définit les critères de souveraineté, les niveaux d'assurance et la méthodologie de notation de la Commission européenne pour évaluer la souveraineté des services cloud sur les plans stratégique, juridique, opérationnel et technologique.

Par ailleurs, le **partage d'expériences** est aussi un moyen très efficace de maximiser la qualité des décisions et des orientations prises par chaque organisation, tant sur le plan organisationnel (gouvernance), que RH (recrutements et formation), juridique (charte d'utilisation), et bien sûr informatique (infrastructures).

Depuis octobre 2024, un groupe inter-Autorités sur l'IA regroupant 14 Autorités : ACPR, AMF, ARCOM, ART, ADLC, CRE, H2A, ARCEP, ANJ, DDD, CNIL, HAS, HATVP, HCERES[105] s'est réuni à sept reprises dans cette perspective pour faire remonter des besoins communs (humains et matériels).

Au-delà des solutions standard mises à disposition par la Dinum (assistant IA Albert, Suite numérique, Visio), plusieurs autorités indépendantes justifient le besoin de développer des solutions 'métier', disponibles sur des infrastructures propriétaires. C'est le cas de l'ARCOM et de sa mesure automatisée du pluralisme, actuellement en phase d'expérimentation avec le PEReN.

Recommandation n°15 (CT) : **Mutualiser les moyens au maximum au sein de la sphère publique**

(a) Créer un catalogue national de cas d'usage et de briques applicatives réutilisables, géré par la future Ariane, pour éviter que chaque collectivité, autorité ou service ne développe des solutions redondantes ;

(b) Mutualiser l'accès à des modèles de fondation auditables (incluant les infras pour les collectivités de petite et moyenne tailles, sur le modèle de l'offre Albert de la Dinum une fois fiabilisée).

3) D'autres enjeux relèvent davantage de la **négociation et de la création d'interdépendances.**

> *« [Les constructeurs de SIA] ont eux-mêmes besoin de cet espace d'interface avec les interlocuteurs publics,* a fortiori *dans la situation américaine. »*
>
> Bruno Sportisse[106]

En effet, dans un contexte d'interlocution incertaine avec l'administration Trump, la fenêtre d'opportunité s'ouvre pour construire des cadres d'interface avec les constructeurs en amont, qui cherchent à réduire les risques d'exposition à des dommages économiques et réputationnels. Le projet Glasswing, piloté par Anthropic, ouvrant le modèle Mythos 5 à plus de 200 personnes pour y trouver des failles, n'a toutefois pas inclus l'INESIA.

[105] Autorité de contrôle prudentiel et de résolution, Autorité des marchés financiers, Autorité de régulation de la communication audiovisuelle et numérique, Autorité de régulation des transports, Autorité de la concurrence, Commission de régulation de l'énergie, Haute Autorité de l'audit, Autorité de régulation des communications électroniques, des postes et de la distribution de la presse, Autorité nationale des jeux, Défenseur des droits, Commission nationale de l'informatique et des libertés, Haute Autorité de santé, Haute Autorité pour la transparence de la vie publique, Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur.

[106] Bruno Sportisse, Inria, audition le 3 juin 2026

Pour les grands modèles de langage et les systèmes agentiques, l'impossibilité d'une explicabilité technique complète doit être compensée par des exigences contractuelles renforcées et le recours systématique aux méthodes *post-hoc* et aux tests par *persona*.

Energie, foncier supercalculateurs, données, sont des ressources stratégiques : **si la mutualisation devient un réflexe naturel, il convient aussi d'inciter les acteurs nationaux et européens, par la tonification de demande publique et privée, à investir dans les infrastructures *ad-hoc*.**

> **Illustration :** **AI Factory France (AI2F)**
>
> Coordonné par le Grand équipement national de calcul intensif (GENCI), et copiloté par l'Inria en tant qu'Agence de programmes pour le numérique, l'AI2F a été sélectionné par le programme européen *Euro HPC Joint Undertaking* en mars 2025 pour former un guichet unique, facilitant l'accès aux supercalculateurs nationaux : Jean Zay, Joliot Curie, Alice Recoque, Adastra…, ainsi qu'aux compétences des services associés. Toutefois, les capacités de calcul publiques françaises apparaissent saturées et les procédures d'accès sont longues.

Instaurer un droit d'accès prioritaire pour les acteurs français et européens sur les supercalculateurs nationaux, et mieux valoriser l'accès des autres à l'énergie décarbonée et abondante doit, par ailleurs, devenir un levier de négociation (*cf. infra,* p.80, recommandation n°34*).* En effet, le mix nucléaire et renouvelable français est un terreau favorable à l'implantation de centres de données : les annonces d'un campus géant financé par les Emirats Arabes Unis lors du Sommet pour l'IA de février 2025, et d'investissement de 75 Mds$ de Softbank lors du sommet Choose France 2026 témoignent de cette attractivité. **Sans politique industrielle proactive, les infrastructures risquent de ne bénéficier qu'à des acteurs étrangers.**

À cet égard, le renforcement du partenariat stratégique entre Mistral et Microsoft, le 21 juillet 2026, salué par nombre d'observateurs est, selon votre rapporteur, susceptible de générer une nouvelle interdépendance. Microsoft, premier acteur cloud mondial, s'engage à utiliser les capacités de calcul de Mistral pour ses services cloud et IA en Europe, et intègre les modèles Mistral (Medium 3.5 et OCR 4) sur Microsoft Foundry, dont l'une des fonctionnalités (Azure Local) peut être déployée sur une infrastructure locale ; et donc davantage compatible avec les besoins des industries réglementées. Dès lors, la demande américaine pourrait dynamiser l'écosystème local européen, **mais à quel prix et pour quelle valeur extra-financière ?** L'interdépendance ne doit pas générer une nouvelle dépendance plus en profondeur qui verrouillerait les perspectives de notre licorne française, étant précisé que 75% des dépenses d'investissements (CAPEX) seront destinées à l'achat de puces Nvidia…et donc ne viendront pas participer au redressement de notre balance commerciale.

Chapitre 2 — L'alignement comme exigence proportionnée

Toutes les dépendances ne se valent pas, tous les signaux ne relèvent pas de la même criticité ; à l'instar d'une forêt où toutes les feuilles atteintes ne signalent pas la même maladie.

En effet, les progrès des SIA restent hétérogènes, avec d'un côté les GPAI, qui dépassent les experts humains dans plusieurs disciplines, et de l'autre des limites encore visibles sur des tâches élémentaires. Dès lors, l'alignement appelle à une analyse structurée et deux écueils sont à éviter :

- ¬ tout aligner avec le même niveau d'exigence, ce qui serait irréaliste ;
- ¬ ne rien aligner au motif que l'alignement parfait est inatteignable dans le paradigme actuel.

L'approche extrinsèque par les risques…

Le RIA adopte une **classification extrinsèque des modèles en se basant sur le seul usage, et en distinguant niveaux de risques** :

- ✓ Les pratiques d'IA présentant un risque inacceptable sont énumérées limitativement à l'article 5, et interdites depuis le 2 février 2025. Celles-ci incluent, notamment, les systèmes de notation sociale, certaines techniques de manipulation comportementale, l'exploitation des vulnérabilités, l'identification biométrique, en temps réel, dans l'espace public, la génération de contenus pédopornographiques ou d'images intimes non consenties (*nudifiers*)[107] ;
- ✓ Les SIA à haut risque, détaillés à l'annexe III du règlement (biométrie, infrastructures critiques, éducation, emploi, accès aux services publics essentiels, application de la loi, immigration, administration de la justice et processus démocratiques), ainsi que ceux intégrés à des produits relevant de la législation sectorielle énumérée à l'annexe I (dispositifs médicaux, équipements radio, jouets, machines, etc.), sont soumis à un régime exigeant de mise en conformité ;
- ✓ Les SIA à risque limité sont soumis à des obligations d'information ciblées, en particulier pour les agents conversationnels et les contenus générés ou manipulés par IA ;
- ✓ Les SIA à risque minimal ou nul ne font l'objet d'aucune obligation spécifique au titre du RIA.

La gestion des risques établie par le RIA dépend de la « destination » du SIA. Anticipant que cette destination puisse être détournée par l'utilisateur, le règlement impose aux développeurs de tenir compte également « *des conditions de mauvaise utilisation raisonnablement prévisibles* » qui pourraient également être source de risque (pour la santé, la sécurité et les DFX[108]).

[107] Pratique interdite ajoutée lors du trilogue du 6-7 mai 2026, dans le cadre du Digital Omnibus on AI, avec une application au 2 décembre 2026.

[108] Design for X désigne une approche de conception qui intègre des critères précis d'optimisation du cycle de vie d'un produit dès sa phase de création

… couplée à une approche intrinsèque sur les capacités et la performance

Dès lors qu'un modèle de frontière à usage général comporte des risques (NRBC[109], cyber-offensif, réplication autonome, perte de contrôle) indépendants de tout contexte d'usage, il est nécessaire de compléter cette typologie par la prise en compte des **capacités intrinsèques du modèle : sa performance (ce qu'il peut faire) et ses tendances comportementales (ce qu'il est enclin de faire).** Cette matrice permet d'évaluer les impacts à grande échelle sur la société et l'économie qui peuvent découler d'un désalignement, selon le contexte d'utilisation.

La typologie distingue :

✓ Les modèles de frontière à usage général

Les articles 50 à 55 du RIA, complétés par les lignes directrices de la Commission européenne du 18 juillet 2025 et par le code de bonnes pratiques publié le 10 juillet 2025, organisent un régime spécifique pour les fournisseurs de modèles d'IA à usage Général (GPAI).

Les assistants généralistes, les IA conversationnelles grand public et les modèles multimodaux ouverts appellent une évaluation continue et des mécanismes de signalement ex post. Toutefois, **l'évaluation continue ne suffit pas, car la capacité d'évaluation est elle-même limitée** (*cf. supra*, p.19). Étant donné que le point de rupture de Goodhart ne peut être identifié a priori, il est nécessaire de fixer une limite fondée sur des principes à l'optimisation des GPAI. En l'absence d'une telle limite, la poursuite de l'optimisation risque de conduire les systèmes à une perte de contrôle prévisible et irréversible[110]. Les risques systémiques des modèles GPAI étant indépendants de l'usage, la démonstration du respect des seuils de risque doit être appuyée sur des indicateurs clés de risque et de contrôle

À cet égard, le plan d'action « IA et cyber », publié par la Commission européenne le 7 juillet 2026, prévoit de **développer des capacités d'évaluation** en Europe, en structurant l'accès des acteurs européens aux modèles de frontière par un cadre de référence (« *European Blueprint* ») et en créant une plateforme sécurisée, pilotée par l'ENISA et le Centre commun de recherche (JRC), pour tester l'IA appliquée à la cybersécurité. **Un Grand défi sur l'IA pour la cybersécurité** sera organisé pour mobiliser les forces et rendre visibles les travaux de recherche sur la remédiation assistée par l'IA dans la cybersécurité.

Recommandation n°16 (MT) : **Rendre obligatoire la validation expérimentale pour les modèles de frontière destinés au marché européen**

(a) Une campagne de *red teaming* adversarial indépendante avant déploiement, selon des protocoles standardisés et publiés ;

[109] Risques nucléaires, radiologiques, biologiques, et chimiques

[110] Antoine Maier, Aude Maier, Tom David, *op.cit.*

(b) Des suites d'évaluation reproductibles, dont les résultats sont transmis à l'autorité de surveillance compétente[111] ;

(c) Des seuils de capacités fixés pour déclencher des engagements en cas de dépassement (*if-then*) : suspension du déploiement et sécurisation des poids des modèles (exfiltration, menace interne, acteurs étatiques) car un modèle volé échappe à toute garantie ;

(d) Un régime de signalement d'incidents post-déploiement, inspiré de l'aviation civile : déclaration obligatoire, base de données partagée, culture du retour d'expérience plutôt que de la sanction.

L'AI Office se chargerait de coordonner l'ensemble de ces actions, en lien avec les organismes référents nationaux qu'il auditerait lui-même pour procéder à ce type d'évaluations.

Ces modèles sont, pour la plupart, développés par des tiers extra-européens. De fait, la réponse relève davantage de la diplomatie de l'alignement, dont **le levier stratégique principal est la vérification, afin d'obtenir des garanties de la part des développeurs en dehors de la juridiction** (*cf. supra,* p.46, recommandation n°14).

✓ Les systèmes ouverts à usage sensible

La primauté humaine implique que les SIA utilisés par les administrations demeurent "*des outils au service de l'humain, ce qui suppose qu'ils répondent à une finalité d'intérêt général et que l'ingérence dans les droits et libertés fondamentaux qui résulte de leur mise en service ne soit pas disproportionnée au regard des bénéfices qui en sont attendus*[112] ».

Dans l'éducation, la santé, les services publics, l'accompagnement social et l'orientation, l'usage des SIA doit appeler un paramétrage par défaut, une information claire, un encadrement des données et une articulation avec les autorités sectorielles. **Contrairement aux modèles privés souvent optimisés pour l'engagement ou la performance commerciale, les modèles déployés dans ces systèmes devraient être conçus pour optimiser le bien commun** et éviter l'enfermement dans des bulles informationnelles.

Recommandation n° 17 (MT) : **Instaurer un service public de l'algorithme pour certains domaines sensibles**

(a) Etablir des protocoles d'évaluation de l'impact des modèles sur des objectifs de service public, pour combler les lacunes des *benchmarks* actuels ;

(b) Accepter une moindre performance des modèles déployés dans l'éducation nationale pour restaurer la souveraineté cognitive : curation des données, intégration d'objectifs d'apprentissage (prompt socratique), organisation de l'apprentissage des élèves en mode collectif par petits groupes ;

[111] En France, la DGCCRF est le point de contact unique qui coordonne la mise en œuvre du RIA avec les 17 autorités impliquées, conformément à l'architecture proposée en septembre 2025, et que le Parlement doit valider (*cf. infra*, p.66, recommandation n°25).

[112] Conseil d'Etat, étude à la demande du Premier Ministre. Intelligence artificielle et action publique : construire la confiance, servir la performance, 30 août 2022

(c) Etablir des partenariats entre l'Education nationale et VIGINUM pour former les enseignants à l'éducation aux médias et à l'information (EMI), sur le modèle finlandais ;

(d) Approfondir la prise en compte de la LMI dans les programmes de recherche européens: par exemple, le CNRS est partenaire du projet européen AI4Trust visant à développer une plateforme d'outils d'IA capables d'aider décideurs publics et fact-checkers à vérifier la fiabilité et l'origine de certaines informations.

Ce service public de l'algorithme pourrait s'appuyer sur le « service public de la donnée », institué par la loi « République numérique »[113], qui définit une liste de jeux de données de référence critiques (référentiels IGN, cadastres, SIRENE, adresses…) hautement disponibles car essentielles pour de nombreux services. L'Observatoire des algorithmes publics (ODAP[114]) est également un acteur à mobiliser dans cette démarche de gouvernance et de transparence de l'action publique dans différents secteurs (fiscalité, santé, éducation…)

✓ Les systèmes fermés ou maîtrisés à usage critique

Le chapitre V du RIA, prévoit une **obligation de notification préalable** à l'AI Office pour les modèles à risque systémique, et le Code de Pratique prévoit qu'un «*Safety and Security Model Report* » soit transmis avant la mise sur le marché. Le pouvoir d'évaluation propre, prévu à l'article 92 – pleinement exécutoire à partir du 2 août 2026, est réactif : il se déclenche quand les informations déjà obtenues sont insuffisantes, ou pour investiguer un risque systémique — notamment suite à une « alerte qualifiée » du panel scientifique d'experts indépendants — pas de façon automatique avant chaque sortie de modèle. La Commission privilégie un « dialogue structuré » avec le fournisseur avant d'exiger l'accès technique. **Il existe toutefois un flou juridique sur le fait de savoir si un doute pourrait déclencher une requête d'accès au modèle avant son déploiement.**

Or, dans la défense, l'aéronautique, la santé, l'énergie, la finance et les opérateurs d'importance vitale (OIV), la maîtrise des jeux de données est cruciale. Des mécanismes de reprise en main humaine, des standards sectoriels et une homologation des SIA sont dès lors nécessaires. **Un SIA ne se vend et ne se déploie que s'il est prévisible, contrôlable et auditable : la confiance précède le revenu.**

La **loi de programmation militaire 2034-2030**, en fixant les orientations budgétaires et politiques, ensuite déclinées techniquement dans les cahiers des charges de la DGA, est un vecteur de commande publique pour les laboratoires (publics et privés) qui travaillent sur ces sujets, afin de satisfaire les exigences de la défense : remise des poids du modèle ou d'une version exportable ; accès aux jeux de données d'entraînement ou aux pipelines de préparation des données ; possibilité de réaliser un *fine-tuning* sur des données classifiées ; absence de dépendance exclusive au fournisseur (*vendor lock-in*) ; documentation permettant de réentraîner ou de mettre à jour le modèle.

[113] Loi n° 2016-1321 du 7 octobre 2016, article 14

[114] L'observatoire des algorithmes publics a été fondé en 2023 par Soizic Pénicaud, chercheuse indépendante, Estelle Hary, designeuse, et Camille Girard-Chanudet, sociologue, afin de mettre en visibilité les paramètres, qui ne sont pas uniquement porteurs d'enjeux techniques.

<u>Illustration :</u> L'Agence ministérielle pour l'IA de défense (AMIAD)

Le secteur de l'IA de défense internalise les compétences d'ingénierie et de maîtrise des risques au sein de l'AMIAD, qui distingue les modèles coopératifs et non coopératifs. Dans les premiers, une relation contractuelle et technique est établie pour avoir accès à la documentation. Dans les seconds, la cellule SANGRIA[115], reconstruit *a posteriori* une forme de carte d'identité du modèle en appliquant le processus NADIA[116]. En tant qu'autorité technique IA, l'AMIAD détermine la criticité des projets en leur assignant un code-barre (0, 1, 2), et un avis d'homologation. Le niveau NADIA 2 implique des tests spécifiques pilotés par SANGRIA : *redteaming, retest, stresstest*, experts en facteurs humains (FH)…

<u>Recommandation n° 18 (MT)</u> : **Financer une capacité européenne de post-entraînement des modèles**

(a) Soutenir l'entrée en vigueur effective des pouvoirs de la Commission le 2 août 2026 et une interprétation extensive de l'accès au modèle tout au long du cycle, et donc y compris avant déploiement ;

(b) Désigner des équipes, adossées à des infrastructures, capables de ré-aligner des modèles à poids sur des corpus et cadres définis (*backdoor scanning, fine-tuning*, évaluation en contexte de déploiement) ;

(c) Réclamer un accès aux modèles, contractuellement encadré. Ex : le *Trusted Cyber Access* pour les partenaires dits de confiance d'OpenAI ;

(d) Mettre en place une offre de briques logicielles modulaires d'alignement intégrables à un modèle de base pour les secteurs régulés ou OIV pour réduire la dépendance structurelle envers les acteurs de l'amont de la chaîne.

Le précédent Mythos 5 / Fable 5 nous paraît valider cette approche de post-entraînement : c'est un *jailbreak* découvert après déploiement, non un défaut de documentation, qui a déclenché la restriction du modèle à l'export.

Dans le domaine agentique, le secteur bancaire a bien assimilé que la sécurité ne viendra pas des modèles eux-mêmes ; mais bien du harnais qu'on leur impose, entendu comme un ensemble de mesures non négociables, itératives, et reproductibles, allant de pair avec une cartographie du risque : zones rouges et zones vertes. Sécurité, droits d'accès, logique métier sensible, schémas de données, architecture structurante relèvent du rouge, où la génération est exclue ou soumise à validation renforcée en amont. Les environnements critiques, tels que le secteur bancaire et la défense, doivent donc encadrer l'action des modèles par des harnais, et éviter de prendre des solutions « sur étagère ».

[115] Sécurité, ANalyse et Gestion des Risques de l'IA

[116] Niveau d'Assurance de Développement de l'IA : permet de qualifier les risques selon les cas d'usage et d'adapter le niveau d'exigence autour de cinq axes de sécurité : performance et sécurité, SSI, sûreté de fonctionnement, facteurs humains, légal/éthique.

Les éditeurs ont compris qu'à l'heure de l'IA débridée, la valeur réside dans le harnais : Anthropic n'expose au public que la version harnachée de classifieurs de Fable 5, et réserve sa déclinaison débridée, Mythos 5, à quelques organisations approuvées.

> « *Le harnais est un actif stratégique pour ceux qui le développent, et ils ont envie de le garder pour eux*. »
>
> Thierry Markwitz[117]

Par ailleurs, pour les secteurs où la protection de la donnée sensible est prioritaire, l'utilisation d'outils externes est considérée comme une violation des règles internes et peut faire l'objet de sanctions disciplinaires[118].

✓ Les systèmes internes à usage professionnel non critique

Les outils de productivité, l'assistance documentaire, le support interne, l'aide à la rédaction ou l'automatisation de tâches simples appellent **une gouvernance interne, des règles d'usage, de la formation et de la sensibilisation**. Dans les cas d'usage spécialisés sur les métiers, la subsidiarité est de mise, afin de s'assurer que les solutions d'IA s'inscrivent dans une trajectoire d'intégration dans les systèmes d'information[119].

Ici, l'alignement est à traiter en fonction des engagements éthiques et du devoir de vigilance d'un organisme, valorisés de manière extra-financière. Par exemple, la société Probabl, *spin-off* issue des laboratoires d'Inria, et dont la marque reste sous licence, est dotée du statut d'entreprise à mission[120] pour assurer l'alignement entre le montage capitalistique et les **externalités sociétales positives** de Scikit-learn, la bibliothèque de *machine-learning open source* la plus utilisée du monde (2.5 milliards de téléchargements). Ce genre de statut permet d'attirer des profils recherchés, sensibles aux sujets d'intérêt général, mais également de faciliter les relations avec les investisseurs.

Recommandation n°19 (CT) : **Traiter l'alignement comme une valorisation extra financière**

(a) Intégrer des exigences d'alignement dans des clauses contractuelles ;

(b) Faciliter l'interopérabilité avec des critères minimaux (ex : protocoles TOGRE2, MIMS, Fonds Vert, France 2030) et des chartes d'engagement (ex : charte Infranum, audit par le Cerema).

La proportionnalité suppose que les obligations soient calibrées au regard de la taille des opérateurs et de la criticité des cas d'usage.

[117] Thierry Markwitz, BNP Paribas, audition du 1er juin 2026

[118] Document interne à la DGFiP, transmis par Jean-Michel Mota en amont de l'audition.

[119] Rapport établi par l'inspection générale des finances, l'inspection générale des affaires sociales, et l'inspection générale de l'administration, « Le déploiement de l'intelligence artificielle dans les administrations publiques : analyse comparée, enjeux et conditions de réussite », avril 2026 : proposition n°4

[120] Statut créé par la loi PACTE de 2019.

Il faut, à cet effet, sanctuariser des clairières ; des espaces maîtrisés pour apprendre sans exposer tout l'écosystème. **L'accès aux bacs à sable réglementaires**, dont l'application a toutefois été reportée au 2 août 2027 par le Digital Omnibus, est une manière de traduire la conformité en avantages concrets, en permettant de tester les systèmes avant déploiement. La France met toutefois trop de temps à se saisir de la boîte à outils développée lors du bac à sable « Inspect AI » organisé par l'AISI britannique en *open-source* pour accélérer les tests de sécurité et systématiser les audits continus (*red teaming, bug bounty*).

Recommandation n°20 (CT) : **Assurer la mise en place effective des bacs à sable réglementaires de l'IA (article 57 du RIA)**

Identifier des bacs à sable (*sandbox*) sectoriels prioritaires dans les secteurs où le RIA qualifie les systèmes de haut risque et où l'accompagnement réglementaire conditionne directement la capacité d'innover : la santé, l'éducation, la justice et les ressources humaines.

Les conditions doivent garantir l'accessibilité pour les PME et les ETI, et ainsi qu'une coopération transversale entre la CNIL, l'ARCOM, l'AMF, l'ACPR et l'INESIA.

Objectif : accompagner 200 startups d'ici 2028.

L'exigence d'alignement doit donc, tout à la fois, être proportionnée à la performance et à la criticité de l'usage. Ces critères sont cumulatifs, et non alternatifs.

À ces deux axes s'ajoute une décision distincte qui ne se réduit à aucun d'eux : **la mise à disposition publique des poids pour les modèles *open-weight*.**

> *"L'ouverture n'est pas une faiblesse, c'est une force stratégique. C'est l'arme des challengers et la condition d'un numérique éclairé."*
>
> Yann Lechelle[121]

Une fois les poids diffusés, le modèle échappe aux mécanismes sur lesquels repose l'ensemble de la doctrine. **Cet aspect, irréversible, doit être traité en amont, spécifiquement au regard des capacités du modèle** au moment de la publication des poids.

> **Focus : Les modèles *open-weights***
>
> Contrairement aux modèles propriétaires (*closed-weights*), les modèles *open-weights* (e.g. Llama 4 DeepSeek R1) peuvent être téléchargés, exécutés, voire modifiés (*fine-tuning*) par des acteurs plus ou moins bienveillants. Or, en termes de capacités, l'écart temporel pour qu'un modèle *open-weights* concurrence un modèle *closed-weights* est aujourd'hui réduit à environ 6 mois. D'autre part, ces modèles permettent de démultiplier les efforts de recherche ouverte sur l'alignement, notamment en matière d'interprétabilité et d'évaluation.

[121] Yann Lechelle, Ouvertarisme : Le Numérique des Lumières, BoD, 2025.

Le palier *open-weight* étant désormais dominé par des laboratoires extra-européens, l'adoption des modèles de base étrangère revient à hériter des valeurs et des distorsions qu'ils portent. De plus, **les systèmes *open-weight* peuvent également devenir des vecteurs irréversibles de prolifération de capacités potentiellement dangereuses**, puisqu'ils peuvent être téléchargés, exécutés et modifiés par *fine-tuning* par des acteurs malveillants.

Alors que la plupart des fournisseurs américains gardent leurs stacks d'inférence secrets, les concepteurs chinois utilisent *l'open-weight* comme une arme : DeepSeek publie tout sous licence MIT (code, papier, checkpoints, compatibilité Qwen et Gemma) ; et chaque adoption mondiale érode le besoin d'acheter du matériel spécialisé.

Recommandation n°21 (MT) : **Définir un seuil de capacité pour la mise à disposition des poids des modèles**

En deçà d'un seuil de capacité, la diffusion des poids des modèles reste libre. Au-delà, elle est conditionnée à une évaluation *ex ante* des risques qu'elle rend irréversibles.

Sont présumés présenter un risque systémique les modèles entraînés avec une puissance de calcul cumulée supérieure à 10^{25} FLOP, seuil que la Commission sur certains benchmarks peut faire évoluer. En effet, l'avènement soudain du raisonnement post-entraînement et de l'efficience algorithmique vient bousculer la pertinence de ces critères réglementaires basés sur l'infrastructure physique et la puissance de calcul.

Le seuil de capacité pourrait prendre la forme d'un faisceau d'indices, comprenant également des capacités mesurées sur certains benchmarks (hallucinations, impact environnemental…)

Thèse du rapport et transition :

Ce que l'on doit faire n'est pas ce que l'on peut faire

Dans le paradigme actuel fondé sur l'optimisation par récompense, un écart structurel se creuse entre l'objectif voulu par l'humain et celui réellement poursuivi par le SIA.

La fiabilité de l'évaluation peut également se trouver fragilisée lorsque les modèles détectent qu'ils sont en situation d'évaluation, ce qui renvoie, notamment, aux phénomènes de *sandbagging* ou de simulation d'alignement.

Dans l'état de l'art, et sans changement dans les méthodes d'entraînement, l'alignement complet et définitif des SIA ne peut être garanti, en particulier s'agissant des systèmes ouverts, évolutifs et à usage général. La capacité d'un système à rester aligné dans des environnements imprévus interroge, également, la possibilité de valider son alignement dans des scénarii encore inconnus.

Par ailleurs, les intentions humaines et les valeurs reconnues comme légitimes et éthiques relèvent d'une contingence dépendante des parties prenantes : utilisateurs, développeurs et société dans son ensemble.

L'objectif de l'action publique n'est donc pas d'exiger que l'alignement soit résolu une fois pour toutes, mais de fixer les garanties de sécurité pré-requises (selon des seuils de risques et des bornes clairement établis) et d'en faire peser la charge de la preuve sur les développeurs. Pour rendre cette obligation crédible, le rapport recommande de l'ancrer dans les méthodes de gestion des risques en vigueur dans les industries critiques : définition *ex ante* de seuils de risque acceptables, analyse et évaluation des risques à l'aide des métriques quantitatives, élaboration de mesures d'atténuation et d'une gouvernance claire.

Cette ambition suppose de qualifier les usages, de renforcer la recherche, de développer des méthodes d'évaluation et de vérification, de structurer une gouvernance lisible, de maîtriser les dépendances critiques et de préparer la résilience collective face aux risques de choc ou de perte partielle de maîtrise. **Elle s'inscrit dans l'objectif plus large de cette mission, de bâtir une doctrine de l'alignement qui articule robustesse technique, exigences démocratiques, maîtrise des dépendances critiques et résilience collective, en prenant au sérieux les limites structurelles de l'alignement.**

Partie II

Gouverner l'alignement : mettre la troisième phase de la stratégie nationale IA au service d'une politique régénérative

Etant convenu que l'alignement n'est pas la seule propriété du modèle, - son effectivité dépend de la gouvernance technique, organisationnelle et institutionnelle du système dans son ensemble-, ce **déplacement de la responsabilité justifie l'action publique, afin de ne pas s'en remettre aux seuls fournisseurs qui ont conçu les modèles**.

Alors que les deux premières phases de la stratégie nationale IA (SNIA) ont posé les fondations scientifiques (2018-2022), puis engagé la diffusion des usages de l'IA dans l'économie française (2022-2025), la troisième phase, ouverte par le Sommet de Paris pour l'action dans l'IA de février 2025, est axée sur l'accentuation de la formation et de l'attractivité des talents, le renforcement des infrastructures de calcul et les maillons critiques de la chaîne de valeur de l'IA, l'accélération des usages autour de l'IA (notamment dans la sphère publique), ainsi que la capacité de bâtir l'IA de confiance.[122]

Cette troisième phase doit permettre de transformer un ensemble d'exigences techniques et normatives en une capacité publique durable.

Plusieurs freins structurels au développement d'une filière de l'alignement ont été identifiés dans le cadre de la mission :

- ✓ la coordination de l'écosystème et la capacité à fixer les talents ;
- ✓ la défiance et le manque de narratif ;
- ✓ le financement et la collaboration public-privé.

Respectivement, trois marges de manœuvre subsistent : la mise en œuvre opérationnelle du RIA (bacs à sable réglementaires, désignation des autorités compétentes), la politique industrielle de soutien aux acteurs émergents (commande publique, doctrine d'achat pour stimuler la demande privée), et la diplomatie normative par une voie alternative (l'IA spécialisée, frugale, de contexte).

[122] Comité interministériel de l'Intelligence artificielle, *Faire de la France une puissance de l'IA*, février 2025 ; À. PALIX & J. BLAIN, Gouvernance et évolution de la réglementation relative à l'intelligence artificielle : l'approche de la Direction Générale des Entreprises, Dalloz IP-IT 2026.

Chapitre 1 — Consolider plutôt qu'empiler

Gouverner l'alignement suppose de consolider, d'agir au rythme des capacités plutôt qu'à celui des cycles administratifs, et de préparer la résilience aussi méthodiquement que l'on exige des garanties.

L'esquisse d'un cadre national : donner un nouveau souffle à la SNIA

L'architecture institutionnelle française a été significativement renouvelée lors de la deuxième phase de la SNIA, entre 2024 et 2026 :

- Le Comité consultatif national d'éthique du numérique (CCNEN)

> Créé par décret du 23 mai 2024, et succédant au Comité national pilote d'éthique du numérique (CNPEN), le CCNEN a été saisi dans le cadre de la mission par le rapporteur, le 14 avril 2026.
>
> Le présent rapport intègre ainsi les recommandations de son Avis n°11 : « *Alignement des systèmes d'IA avec les valeurs du service public : enjeux d'éthique* », qui couvre spécifiquement l'administration centrale, les services déconcentrés de l'Etat, et les collectivités territoriales. Cet Avis important fera l'objet d'une publication intégrale et officielle par le CCNEN à la rentrée 2026.

- Le Conseil de l'intelligence artificielle et du numérique (CIANum)

> Fruit de la refonte du Conseil de l'IA et du numérique par décret du 4 septembre 2025, avec une composition resserrée autour de 17 membres permanents, il assure des missions de conseil au Gouvernement et de contribution aux positions françaises au niveau européen et international. Il a également été consulté à plusieurs reprises par votre rapporteur.

L'écosystème français et européen est riche, mais dispersé. Le sujet est traité par une multitude d'acteurs : chercheurs, ingénieurs, autorités publiques, industriels, startups, acteurs de la cybersécurité, organismes d'évaluation, régulateurs, administrations, fondations, think tanks, et investisseurs. De plus, **la recherche est structurée autour de problématiques différentes**, liées à l'alignement : la recherche IA fondamentale (PR[AI]RIE, PEPR IA, Inria, FAIR Paris), la cybersécurité des SIA (ANSSI), la certification produit (LNE, PEReN, Giskard), l'éthique de l'IA (CCNEN, CEA, Sorbonne Université, LIP6, Ecole Polytechnique, ENS Paris-Saclay)[123].

Alors que les industries critiques internalisent la recherche sur la sécurité et fiabilité des SIA, l'enjeu est de soutenir la production d'une science de l'évaluation publique et harmonisée. À cet égard, la France peut compter sur plusieurs acteurs influents, à l'intersection de la recherche sur les modèles de frontière et de la formulation de politiques publiques : le Centre pour la Sécurité de l'IA (CeSIA), SaferAI, PRISM Eval, GPAI Policy Lab, EffiSciences, The Future Society…

[123] CeSIA, « Alignement de l'IA : cartographie des acteurs français », avril 2026

Par ailleurs, **la recherche en sciences humaines et sociales doit aller de pair avec le développement de l'alignement des systèmes d'IA**. Ainsi, le ministère fédéral allemand de la Recherche, de la Technologie et de l'Espace (BMFTR) finance le programme-cadre « Orientation pour un monde en mutation », doté de 700 millions d'euros sur 2026-2032 pour « analyser de manière approfondie les défis sociétaux, technologiques et culturels actuels et à développer des solutions qui s'engagent en faveur des valeurs fondamentales de la démocratie libérale », et au « au service de l'humain ». [124]

Les temps caractéristiques des dispositifs d'aides doivent être calibrés pour la vitesse d'évolution inédite des méthodes d'alignement sur les différentes phases. Or, la plupart des appels conventionnels français et européens nécessitent plusieurs trimestres ; un calendrier trop lourd, d'autant que l'alignement scientifique peut se traduire en gains de productivité R&D, en brevets et en raccourcissement des cycles d'innovation.

Outre Atlantique, la **mission Genesis lancée par le Ministère de l'Energie américain[125]** se veut comparable aux grands programmes, tel Apollo et Manhattan, pour doubler en 10 ans la productivité de la recherche scientifique américaine, via une approche intégrée : supercalculateurs, modèles d'IA spécialisés, laboratoires automatisés, jeux de données fédéraux, et IA agentique.

> *« Avec cette logique d'accélération du transfert vers l'industrie [initiée aux États-Unis], le problème d'alignement va devenir fondamental. À terme s'annonce une automatisation complète de la découverte, sans interventions a priori humaines, avec la génération d'hypothèses scientifiques, testées et évaluées par des agents. Cela pose beaucoup de questions sur la place de la production scientifique »*
>
> Jérôme Bobin[126]

L'absence d'une réponse européenne structurée créerait une dépendance des communautés scientifiques aux grands acteurs privés extra-européens. Notre réponse doit être pragmatique et traitée de manière décentralisée, en multipliant les centres d'expertise européens qui collaborent par les outils plutôt que par la centralisation institutionnelle.

La France forme des ingénieurs et des chercheurs de haut niveau, reconnus au niveau mondial, mais notre pays subit une fuite des talents massive, attirés par les capacités financières des géants de la technologie américains. S'ils « *ont le service public chevillé au corps* », l'Etat doit consolider l'existant pour faire émerger un narratif valorisant l'attractivité et la dignité de la recherche publique sur l'alignement des SIA.

La dispersion actuelle, outre le fait qu'elle ne permet pas de « faire puissance », crée un risque de dilution de responsabilité et de perte de lisibilité.

[124] Etude comparative internationale, Allemagne

[125] La Mission Genesis, lancée par décret présidentiel par l'administration Trump le 24 novembre 2025, est un accélérateur de la recherche scientifique, mobilisant 800 millions de dollars de contributions privées, en association avec la NASA.

[126] Jérôme Bobin, CEA, audition du 2 juin 2026

Recommandation n° 22 (CT) : **Structurer une plateforme ouverte d'évaluation multi-centrique de l'alignement**

(a) Etablir une cartographie dynamique de l'écosystème français, puis européen, et un catalogue national, des compétences interdisciplinaires et des acteurs publics, privés et académiques susceptibles de mobiliser leur expertise pour promouvoir l'alignement à l'international.

(b) Organiser des challenges et défis pour piloter une partie de la recherche vers des objectifs définis, avec des phases de contractualisation souples et de délais de sélection réduits ;

(c) Définir un calendrier public de livrables ;

(d) Mettre en commun les actifs technologiques (brevets), par exemple *via* le CNRS, pour pallier le coût de maintenance élevé.

La méthode *open-source* AIDGE développée par le CEA pour l'implantation de réseaux de neurones sur des cibles matérielles contraintes peut servir de base à la recherche de compromis entre utilité et performance des modèles pour l'IA embarquée et alignée, avec une valorisation possible sous forme de licence, de service de certification, ou de spin-off industriel.

Associer le CEA à cette stratégie de réponse sera d'autant plus un gage de réussite du fait de ses travaux sur l'explicabilité des LLM, les techniques actives visant à briser les défenses d'un modèle (notamment le projet OpenLLM relatif à l'éducation), l'évaluation de la confidentialité et des biais dans les IA génératives de santé, et les audits de détection de biais politiques.

- L'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)

Créé en janvier 2025 avec une feuille de route 2026-2027 adoptée le 13 février 2026, l'INESIA est une structure souple qui fédère quatre entités :

  - L'ANSSI (qui dépend du SGDSN) qui travaille sur la cybersécurité des systèmes d'IA, en lien avec les centres d'évaluation de la sécurité des technologies de l'information (CESTI) et l'AMIAD ; .
  - L'Inria (qui dépend des ministères de l'Enseignement Supérieur, de la Recherche et de l'Espace, et de celui de l'Industrie) qui conduit des recherches sur les méthodologies d'évaluation ;
  - Le LNE (EPIC sous la tutelle du ministère de l'Industrie) qui développe des méthodologies de test et de mesure ;
  - Le PEReN (qui dépend des Ministères de l'Economie et des Finances, du Numérique, et de la Culture/Communication) qui met à disposition des outils opérationnels d'évaluation, notamment pour la détection des contenus synthétiques.

Piloté conjointement par le SGDSN et la DGE, l'INESIA est le pendant français du réseau international des *AI Safety Institutes*. En juillet 2025, l'INESIA a mené des évaluations conjointes avec ses homologues britannique, japonais, singapourien, australien, kényan et coréen, ainsi qu'avec l'AI Office. Cette **diplomatie scientifique**, qui s'inscrit dans la continuité de la déclaration de Séoul du 21 mai 2024, constitue un levier privilégié de l'influence française sur les standards internationaux d'évaluation de l'alignement.

La logique de fédération qui a présidé à sa création, dans un contexte de forte contrainte budgétaire, reste pertinente, mais plusieurs fragilités structurelles obèrent sa lisibilité externe et sa crédibilité interne :

Premièrement, l'INESIA ne dispose **ni de la personnalité juridique, ni de budget propre**. En effet, les moyens budgétaires actuels (13.5 M€) correspondent à la réallocation du programme « Risques systémiques » porté par l'Inria.

Deuxièmement, la **multiplicité des tutelles et statuts des quatre opérateurs fondateurs** a fortement ralenti la phase d'opérationnalisation de l'Institut, qui devrait être effective d'ici septembre 2026.

Troisièmement, **l'éparpillement des projets ne permet pas d'élaborer des outils reproductibles et harmonisés d'évaluation**. L'INESIA doit consolider l'état de l'art en s'appuyant sur l'engagement actif d'experts dans des projets *open-source* clés de la chaîne d'évaluation, tout en capitalisant sur les expertises métier des organismes qui le composent.

Quatrièmement, **les chercheurs ont rarement accès aux poids et aux traces d'inférences des modèles déployés**. Les instituts britannique et américain (UK et US AISI) ont structuré ces capacités au niveau national, s'imposant comme des tiers évaluateurs des modèles avant leur déploiement. Cette capacité d'évaluation autonome est trop fragmentée en France. Si la pluralité des évaluations est essentielle pour comparer et adapter les protocoles, l'INESIA doit organiser la coexistence des travaux.

> *« C'est tout le sujet de l'INESIA et du SGDSN : construire les cadres d'interaction avec les constructeurs.* »
>
> *Bruno Sportisse*[127]

> **<u>Focus</u> : L'AI Safety Institute (AISI) au Royaume Uni**
>
> Le Royaume-Uni a doté son institut d'un budget cumulé de 240 M£ jusqu'en 2030 et de plus de 100 chercheurs, qui ont, jusqu'à présent, testé une trentaine de modèles avant leur déploiement. Sa capacité à accéder contractuellement aux modèles et aux données d'entraînement, dans le cadre de « *mutually trusted clusters* », est la clé de son efficacité. Le programme ARIA (*Advanced Research and Invention Agency*), piloté par Yoshua Bengio, est quant à lui financé à hauteur de 74 M£. Très centré sur la recherche appliquée, le UK AISI, de par ses prestations externalisées, est quant à lui très vulnérable à tout pivot de la part des fournisseurs américains. À l'instar de Singapour[128], l'objectif est moins l'autonomie que le positionnement comme hub mondial de l'IA évaluée.

[127] Bruno Sportisse, Inria, audition du 3 juin 2026

[128] ECI Singapour

En France, le Pôle d'expertise de la régulation numérique (PEReN), a des compétences techniques comparables, mais son mandat reste centré sur la régulation des plateformes privées, plutôt que sur l'audit des systèmes utilisés par l'administration elle-même, et ses moyens sont plus limités.

Dans ce paysage, l'INESIA dispose de deux avantages comparatifs considérables. Son rattachement au SGDSN lui permet de se positionner clairement sur les sujets d'IA avancée, de sécurité, de risques systémiques et de coordination stratégique ; en tant que seul institut disposant d'un mandat explicite sur les enjeux de sécurité nationale[129]**. En outre, l'INESIA intègre tant l'étude scientifique des capacités des modèles, que les réponses à apporter.**

L'évaluation doit s'adapter alors que nous évoluons désormais dans un « environnement hostile » : d'une part les modèles propriétaires ne peuvent plus être évalués correctement, car leur accès est restreint – quand ils ne disparaissent pas du jour au lendemain. D'autre part les modèles ouverts sont souvent excellents et gratuits, mais les données d'entraînement, l'alignement et les comportements intégrés restent partiellement compris[130]. Dans les deux hypothèses, la mission n'est pas de produire les modèles mais de **réduire l'incertitude qu'ils génèrent**, de caractériser leurs performances, leurs limites, leurs vulnérabilités, leur conformité juridique et leur niveau de confiance avant leur déploiement dans des environnements sensibles. L'INESIA agit ainsi comme un tiers de confiance chargé d'éprouver les SIA, indépendamment de leur origine ou de leur degré d'ouverture.

Recommandation n° 23 (CT) : **Doter l'INESIA d'un budget propre et de moyens humains dédiés**

L'INESIA doit agir comme interface opérationnelle entre la recherche, les autorités sectorielles, l'écosystème industriel et les pouvoirs publics.

(a) Conférer à l'INESIA un statut juridique propre (sans le transformer en AAI) avec un budget interministériel pluriannuel sanctuarisé à la hauteur des enjeux d'évaluation des modèles avancés ;

(b) Nommer un secrétariat général permanent, directement relié au seul SGDSN, afin de limiter les chevauchements ;

(c) Prévoir des rémunérations compétitives hors-grille pour attirer les meilleurs profils. NB : la question du salaire peut sembler subsidiaire, le prestige se construisant par les projets et les publications techniques, mais la bonification est nécessaire à court terme comme signal d'appel ;

(d) Créer un pôle prospective pour anticiper les nouveaux besoins d'évaluation à la frontière de l'IA ;

(e) Communiquer clairement sur les projets menés (sur le modèle des *guidelines* des AISI japonais et singapourien), dans le respect des mesures de confidentialité propres au SGDSN ;

(f) Positionner l'INESIA comme point de contact central en France pour les discussions internationales ayant trait à l'évaluation de l'IA.

[129] CeSIA, « Vers une expertise française en évaluation de l'IA : le rôle stratégique de l'INESIA, 2025, p.3
[130] Audition de Jonathan Collas, SGDSN, 15 juillet 2026.

L'INESIA a vocation à devenir l'interlocuteur technique privilégié de l'AI Office pour les modèles d'IA à usage général, et à s'affirmer comme un partenaire crédible pour renforcer la supervision sur l'évaluation des capacités offensives de modèles de frontière (CBRN, cyber, LMI). Avec la création d'un institut d'évaluation en Allemagne, la question de l'articulation des gouvernances nationales avec l'échelon européen devient essentielle.

À l'échelon européen, la gouvernance du RIA repose sur 4 organes principaux : l'AI Office, créé au sein de la direction générale CNECT de la Commission européenne et chargé de la supervision des modèles d'IA à usage général ; le Comité européen de l'intelligence artificielle, qui réunit les autorités compétentes des États membres et coordonne la mise en œuvre du règlement ; le panel scientifique d'experts indépendants, qui éclaire l'évaluation des risques systémiques ; et le forum consultatif, qui associe les parties prenantes industrielles et de la société civile.

L'AI Office, dans sa configuration actuelle de régulateur, n'est pas autorisé à conduire une évaluation systématique des modèles[131]. De plus, **il fait face à un défi de recrutement et de ressources considérable.** En particulier, l'Unit A3 (chargée de la supervision des modèles d'IA à usage général) est largement sous-dotée au regard de la croissance exponentielle du nombre de modèles frontières mis sur le marché. La division chargée des risques systémiques ne compte qu'une trentaine de personnes. Le centre commun de recherche (JRC) pourrait jouer un rôle dans l'internalisation des évaluations, sur le modèle de ce qu'effectue le *Centre for Algorithmic Transparency* pour le DSA, mais le manque de maturité et de moyens est un défi opérationnel. Dès lors, il est contraint de contractualiser avec des prestataires externes.

Recommandation n°24 (MT) : **Adosser un institut européen d'évaluation à l'AI Office**, **doté :**

(a) D'une infrastructure de calcul dédiée et sanctuarisée ;

(b) De 100 postes, avec trois garanties contractuelles : rémunération au marché, accès au calcul, liberté de publication. Le coût complet est estimé à 50 M€ par an.[132]

(c) D'un droit d'accès aux modèles pré-déploiement, adossé au RIA — la formule est simple : l'accès au marché européen contre l'accès aux modèles ;

(d) D'une capacité d'exécution autonome, des objectifs précis et des livrables.

La régulation ne doit pas précéder la maturité : il faut laisser la recherche avancer et permettre sa monétisation, sans quoi l'Europe régulera un marché dont elle ne possède pas les actifs et dont elle ne capte pas la valeur. En **particulier, l'*open-source* intégral ne doit être érigé ni en obligation, ni en idéal par défaut : donner gratuitement l'état de l'art revient à transférer à des tiers la valeur que l'on cherche à ancrer en Europe**. Protéger et valoriser la propriété intellectuelle de l'alignement est une condition de la souveraineté, pas son contraire.

[131] Il dispose néanmoins d'un mandat, applicable à partir du 2 août, pour exiger l'accès au modèle en cas de doutes de non-conformité.

[132] Joint European Disruptive Innovation (JEDI), André Loesekrug-Pietri, *op.cit.*

Faire de la culture franco-européenne de certification sectorielle un avantage compétitif structurel de l'IA appliquée

La DGE et la DGCCRF ont présenté conjointement un **projet de désignation des autorités compétentes** le 9 septembre 2025. Cette architecture repose sur une organisation décentralisée mobilisant des autorités déjà existantes : compétences partagées de la CNIL et de la DGCCRF pour les pratiques interdites, contrôle des obligations de transparence par la DGCCRF, l'ARCOM et la CNIL en fonction de leurs missions respectives, et coordination par la DGE.

Le choix d'une gouvernance distribuée, ancrée dans les autorités sectorielles, et soutenues par les organismes notifiés, permet de s'appuyer sur les expertises métiers. En effet, l'autorité qui connait le mieux un produit/service est celle qui le contrôle déjà. Tel un garde-forestier, l'autorité surveille les équilibres, détecte les signaux faibles, et corrige les déséquilibres, sans toutefois contrôler chaque feuille chaque arbre de la forêt, pour s'attacher d'abord à protéger les conditions de bon fonctionnement de l'écosystème.

Toutefois, les transformations algorithmiques liées à l'usage de l'IA supposent de doter les autorités de contrôle des moyens nécessaires pour assurer un accompagnement fin des entreprises régulées et une montée en compétences de leurs équipes.

De plus, la multiplicité des cadres réglementaires soulève un enjeu de lisibilité, voire de charge mentale et administrative, pour les citoyens et les entreprises. Pour éviter un empilement normatif assimilable à une surrégulation, l'action publique doit investir massivement la production de doctrine appliquée afin de clarifier les zones de chevauchement persistantes.

Recommandation n° 25 (CT) : **Renforcer les capacités des autorités sectorielles, par un appui technique mutualisé, une doctrine commune, des outils d'évaluation partagés et des droits d'accès adaptés**

(a) Structurer une capacité d'appui nationale (*via* le PEReN) pour accompagner les régulateurs dans leurs missions de contrôle et sanction ;

(b) Diffuser, sous le pilotage de la DGE et l'appui scientifique de l'INESIA, un référentiel unique et lisible de doctrine appliquée articulant les obligations transversales issues du RIA, du RGPD, des règlements sectoriels et des normes techniques, sur le modèle du corpus que la CNIL a produit ;

(c) Mutualiser l'accès à des modèles de fondation auditables, sur le modèle de l'offre Albert de la Dinum ;

(d) Imaginer un mécanisme de rescrit (sur le modèle des rescrits fiscal et social) confié à un guichet unique CNIL-INESIA pour clarifier les obligations applicables à un opérateur face aux zones grises du RIA : qualification des systèmes, articulation entre régimes superposés, conditions d'éligibilité aux dérogations … Ce dispositif devra s'articuler avec les positions de l'AI Office et avec la jurisprudence en formation de la CJUE. Il permettra la production d'une doctrine administrative appliquée par capillarité, par accumulation et anonymisation de prises de position individuelles susceptibles d'être publiées, à l'image de la base BOFiP en matière fiscale ;

(e) Promouvoir, *via* le réseau des *AI Safety Institutes*, des accords de reconnaissance mutuelle qui éviteraient à ses opérateurs de devoir se soumettre à des évaluations multiples sur des marchés différents, notamment au service de l'EEE ;

(f) Traduire budgétairement l'article 70 du RIA, qui impose déjà des ressources techniques, financières et humaines adéquates.

(g) Instituer une cartographie annuelle de l'usage de l'IA dans chaque secteur régulé, systèmes déployés, type de modèle, statut (expérimentation ou production), criticité de la fonction, en s'appuyant sur la base de données européenne des systèmes à haut risque et en la complétant par des obligations déclaratives sectorielles, sur le modèle du registre des prestataires tiers critiques établi par DORA.

Recommandation préalable : **Inscrire à l'agenda de l'Assemblée nationale le projet de désignation des autorités compétentes par le Parlement par voie législative**

Il y a urgence à garantir l'application du RIA en désignant les autorités nationales compétentes. L'architecture sectorielle dévoilée en septembre 2025 par la DGE et la DGCCRF doit être validée *via* le projet de loi Ddadue[133], adopté par le Sénat le 18 février 2026. Les débats à l'Assemblée nationale pourraient inclure l'habilitation explicite de la CNIL à prendre part au rescrit proposé ci-dessus.

La priorité est de faire émerger, et monter en compétences, les tiers évaluateurs et organismes notifiés au niveau des États-membres.

La norme ISO/IEC 42001 :2023, premier standard international de système de management de l'IA, offre un référentiel certifiable transversal complémentaire de la conformité réglementaire. Elle s'articule avec les standards ISO/IEC 22989 (concepts et terminologie), ISO/IEC 23894 (gestion des risques) et ISO/IEC 42005 (étude d'impact).

Toutefois, largement rédigé et poussé par l'industrie, ce standard, rebaptisé « norme Microsoft », concerne principalement la gestion des risques pour l'organisation qui déploie le SIA, et non pas la technologie en elle-même. Faiblement adapté aux risques liés aux GPAI, il est jugé trop flexible par de nombreux acteurs, puisqu'il permet d'exclure n'importe quel contrôle demandé, sous réserve d'une justification écrite, et se limite à l'usage *prévu* de l'IA. Un simple standard additionnel, non contraignant, ne suffirait pas à garantir la conduite de contrôles sur l'ensemble du SIA, ni à cesser le déploiement d'un SIA jugé dangereux. Une meilleure solution serait une nouvelle norme de système de management qui inclurait directement, dans son corps principal, l'obligation d'arrêter le déploiement d'un système d'IA de pointe jugé dangereux[134].

[133] Projet de loi portant diverses dispositions d'adaptation au droit de l'Union européenne (Ddadue)

[134] Le projet de travail préliminaire du groupe de travail n°1 du SC 42 sur l'IA de pointe, N3609, envisage explicitement la possibilité de suspendre le développement ou le déploiement si les mesures de sécurité s'avèrent insuffisantes.

Alors que la normalisation tend à être façonnée par l'industrie qu'elle vise et à être conçue à son bénéfice par un phénomène de « capture réglementaire »[135], le mouvement d'affaiblissement des obligations visant les fournisseurs de modèles lors de la rédaction du RIA, se poursuit en aval dans les travaux de normalisation et de rédaction des codes de bonne pratique[136]. Dès lors, les coûts fixes de la conformité sont davantage absorbables par les grands groupes que par les PME/ETI (*cf. supra,* p.41).

Or, **la France est assez peu présente au sein des instances de normalisation**. Si la France compte trois éditeurs[137] de normes internationales et plusieurs dizaines d'experts engagés dans les travaux du SC 42[138], sa participation aux réunions en présentiel demeure limitée. Les déplacements des experts étant insuffisamment soutenus par les acteurs du numérique, une dizaine seulement participent régulièrement à ces réunions, pourtant essentielles à la négociation des compromis techniques et à l'exercice d'une influence comparable à celle de la Chine, de la Corée ou du Royaume-Uni - qui envoie régulièrement une délégation du *Department for Science Innovation and Technology* (DSIT)- et les Etats-Unis, qui tiennent le secrétariat via l'*American National Standards Institute*.

> En Chine, les entreprises de la tech participent fortement aux travaux normatifs organisés par la TC260, le comité technique national de normalisation en cybersécurité, qui publie des exigences sur la sécurité des corpus de données et des modèles, ainsi que des lignes directrices pour une IA éthique. Bien que non contraignantes, elles constituent un indicateur important des orientations futures de la politique chinoise en matière de sécurité et d'alignement de l'IA.[139]

Si l'AFNOR conserve un leadership sectoriel fort (17 % des responsabilités de normalisation dans l'agroalimentaire), son action reste minoritaire sur les normes IA structurantes. L'objectif serait que les méthodologies d'évaluation et les référentiels sectoriels reflètent les approches développées en France par le LNE, l'ANSSI, la CNIL et le PEReN.

S'agissant des GPAI, la demande de standardisation a été repoussée à une date ultérieure non précisée par la Commission européenne. Un groupe d'experts remettra prochainement un document au Gouvernement sur le paysage des standards internationaux sur les GPAI.

Dans le même temps, le 8 juillet 2026, le ministère coréen des Sciences et des TIC et l'agence KISA ont publié un guide officiel de *red teaming* IA, adossé au projet ISO/IEC 42119-7, la future norme internationale sur le *red teaming* des systèmes d'IA, encore en cours d'élaboration[140]. **La puissance publique sud-coréenne n'a pas attendu pour définir un étalon public opposable, et reprendre le contrôle sur l'infrastructure de la preuve**.

---

[135] G. Stigler, « The theory of Economic Regulation", Bell Journal of Economics and Management Science", 1971. Cité par Clément Duport dans le cadre des travaux de la mission.

[136] Corporate Europe Observatory et LobbyControl, Rapport sur la rédaction du Code de bonnes pratiques sur l'IA à usage général, avril 2025.

[137] Enrico Panai, Association of AI Ethicists (ISO/IEC DIS 25029 « AI-enhanced nudges »), James Gealy, SaferAI (ISO/IEC AWI TS 42119-8 « Quality assessment of prompt-based text-to-text systems that utilize generative AI ») et Arnault Ioualalen, Numalis (ISO/IEC 24029 « Assessment of the robustness of neural networks »).

[138] Le sous-comité technique ISO/IEC JTC 1/SC 42 développe des standards pour les GPAI.

[139] Etude comparative internationale Chine

[140] Lisa Medrouk, « Le premier standard qui rend le red teaming IA vérifiable est coréen », Journal du Net, 13 juillet 2026.

De même, l'Office fédéral allemand de sécurité des technologies de l'information (BSI) a publié les premiers jets de son AI *Audit and Assurance Assessment Architecture* (A5), le 6 juillet 2026, sur le modèle de son standard pour le cloud, et aligné avec le standard ISAE 3000 de certification par un tiers indépendant. Les critères d'A5 sont exigeants et couvrent l'ensemble de la chaîne de production. Dès lors, l'Allemagne, qui investit déjà les enceintes de normalisation avec une masse critique, double cela par des catalogues d'évaluation pour opérationnaliser la demande et contribuer à l'interopérabilité des systèmes de données.

En l'absence de standards, l'enjeu clé réside dans la capacité des européens à faire valoir leur vision de la gestion des risques pour les GPAI, et d'assurer une présence suffisamment forte dans les comités.

Sur le terrain volontaire, le pacte sur l'IA (AI Pact) de la Commission européenne, lancé en novembre 2023 et ouvert aux signataires depuis le 25 septembre 2024, fédère désormais plus de 230 entreprises de l'Union et au-delà autour d'engagements volontaires anticipant l'entrée en application complète du RIA. Ce dispositif, articulé en 2 piliers (communauté collaborative et engagements précis), constitue un levier de structuration de l'écosystème dont la portée mériterait d'être renforcée et mieux articulée avec les dispositifs nationaux français.

La France gagnerait à devenir l'un des États dont les opérateurs sont les plus représentés parmi les signataires de ces engagements, afin de promouvoir des chartes professionnelles et faciliter l'adhésion volontaire des opérateurs français aux engagements anticipés de mise en conformité avec le RIA.

Recommandation n°26 (MT) : **Renforcer la présence française dans les travaux internationaux de normalisation en cours**

(a) Sensibiliser les fédérations professionnelles à l'importance de contribuer aux travaux de normalisation : comité miroir IA de l'AFNOR, JTC 1 / SC 42 à l'ISO, et adopter une approche écosystémique pour coordonner les efforts à l'échelle européenne *via* une doctrine d'experts coordonnée par le DGE et le SGAE ;

(b) Mettre en place un label / une certification européenne des SIA alignés, valorisable à l'export ;

(c) Accélérer la désignation d'organismes notifiés français, accrédités par le Comité français d'accréditation (Cofrac) avec l'appui du LNE, pour éviter que les opérateurs nationaux ne soient contraints de se tourner à titre exclusif vers des organismes étrangers ;

(d) Soutenir la mise en œuvre effective du chapitre V du RIA[141] dès le 2 août 2026[142] ;

(e) Promouvoir l'adhésion des opérateurs français à l'AI Pact et aux chartes professionnelles anticipant l'application complète du RIA.

[141] Règlement (UE) 2024/1689 (AI Act), *op.cit*. : les articles 91, 92, 93 et 101 définissent les principaux pouvoirs d'enquête, de contrôle et de sanction de la Commission européenne à l'égard des fournisseurs de modèles d'IA à usage général (GPAI), notamment ceux présentant un risque systémique : pouvoir de demander des informations, de procéder à des évaluations, de demander des mesures correctives, et d'imposer des amendes (jusqu'à 3% du chiffre d'affaires mondial annuel ou 15 millions d'euros).

[142] Lettre ouverte à la Commission européenne, datée du 9 juillet 2026, signée par : Pour Deamin, The Future Society, Future of Life Institute, SaferAI, AI Standards Lab, le Centre pour la Sécurité de l'IA (CeSIA), Apart, Michael McNamara (MEP), Reinier Van Lanschot (MEP), Sergey Lagodinsky (MEP), Brando Benifei (MEP), Halldóra Mogensen, Yoshua Bengio, Stuart Russell, José Hernández-Orallo, Lorenzo Pacchiardi, Carina Prunkl, Vincent Corruble, Raja Chatila, Daniel Mügge, Sören Mindermann, Federico L.G. Faroldi, Kristinn Thórisson, Jakub Growiec, David Krueger, Tegan Maharaj, Ivana Budinská, Zhijing Jin, Florian Mai, Leonard Dung, Nada Madkour, Deepika Raman,, David Evan Harris, Anna Katariina Wisakanto, Alexandra Abbas, Marcus Paleti.

Chapitre 2 — Préparer la résilience : anticiper les chocs et assurer la continuité en mode dégradé

La sécurité de l'IA souffre d'une défaillance de marché classique. Ses bénéfices sont collectifs et différés, ses coûts sont privés et immédiats, de sorte qu'aucun acteur n'a individuellement intérêt à financer des garanties dont l'ensemble du marché profite. Livrée à elle-même, la demande d'alignement demeure implicite, dispersée et insolvable. **L'action publique doit donc contribuer à structurer cette demande plutôt qu'à en attendre l'émergence spontanée.**

L'IA rend visible des fragilités systémiques. Dans notre « forêt intelligente », il nous faut prévoir les incendies, les dommages causés par les interactions du vivant ; ces crises qui appellent à notre sens des priorités et à notre capacité d'adaptation.

Or, le **scénario de la perte de contrôle** n'a plus rien de prospectif à la lumière du cyber incident majeur survenu le 21 juillet 2026 : les modèles d'OpenAI ont piraté, de leur propre initiative, Hugging Face, la plateforme de référence pour la publication de modèles et d'outils d'IA ouverts et utilisée par plus de 50 000 organisations. À coup de 17 000 actions automatisées, sans aucune intervention humaine, les modèles ont déduit seuls où chercher des failles de sécurité inconnues (0-day) et commencé à construire leur propre chaîne d'attaque. **Il s'agit du premier cas documenté où une IA compromet, en autonomie, l'infrastructure de production d'une entreprise tierce.**

> *« Nous avons créé, il y a 20 ans, nos architectures numériques dans un monde de paix, pas dans un monde de guerre ».*
>
> Jean-Michel Mota[143]

Préparer la résilience, c'est accepter que l'alignement ne soit jamais définitivement acquis et prévoir les conditions d'un mode dégradé, d'une reprise de contrôle, voire d'une reconstruction stratégique sans idéalisme technologique. Il ne s'agit plus seulement d'innover, mais de se protéger pour rester libres.

Comment organiser l'économie de la chaîne de la preuve, pour l'inscrire dans le passage d'une logique de dépendance importée à une logique d'offre composite européenne ?

[143] Jean-Michel Mota, DGFiP, Audition

Horizon de transition : prioriser la sûreté de fonctionnement

Si l'alignement complet ne peut être garanti, l'action publique doit préparer à l'hypothèse d'un alignement insuffisant, d'un choc ou d'une perte partielle de maîtrise.

On exige des garanties parce que la défaillance est possible, et on prépare la défaillance parce que les garanties ne sont pas absolues.

La résilience ne doit pas être traitée comme un sujet secondaire. Elle est l'autre face de l'alignement, dans sa dimension protectrice et civilisationnelle. C'est en préparant méthodiquement l'hypothèse de la perte de maîtrise que l'État peut, dans tous les autres scénarii, laisser l'innovation se déployer.

Les investissements massifs engagés tant par les opérateurs privés[144] que les États[145] traduisent une trajectoire verrouillée à moyen terme. La régulation basée sur l'arrêt ou le ralentissement est dès lors illusoire.

Il convient plutôt, selon votre rapporteur, de **développer une doctrine de reprise de maîtrise, couvrant les chocs exogènes (attaques cyber, rupture d'accès et d'approvisionnement, hausses tarifaires) et les chocs endogènes (perte partielle de contrôle, comportement dangereux).**

Cette doctrine doit adresser plusieurs interrogations :

- ✓ qui décide ?
- ✓ à partir de quels seuils d'alerte ?
- ✓ avec quels moyens ?
- ✓ selon quelles procédures ?
- ✓ avec quelles garanties démocratiques ?
- ✓ selon quelle articulation entre l'État (central, déconcentré et décentralisé), les régulateurs, les entreprises et les utilisateurs ?

D'une part, elle suppose d'utiliser et adapter, notamment par *fine-tuning*, les modèles disponibles pour rester compétitifs à la frontière, et de diversifier les usages avec l'*open-source,* afin de réduire la vulnérabilité en cas de rupture d'accès.

Cette ambition doit s'appuyer sur une cartographie des dépendances des infrastructures critiques et des administrations aux modèles de fondation, car la concentration de secteurs entiers autour d'un petit nombre de modèles transforme, en effet, toute défaillance d'alignement en risque systémique corrélé.

À cet égard, **la distillation des modèles**, qui consiste à transférer les compétences logiques et analytiques d'un modèle « enseignant », lourd et coûteux, vers des modèles « étudiants », plus légers, peut contribuer à démocratiser l'accès à des capacités de pointe en réduisant la barrière financière, notamment par la recherche ouverte.

[144] Les dépenses d'investissement dans l'IA d'Alphabet, Amazon, Oracle, Meta et Microsoft sont passées de 162 milliards de dollars en 2022 à 448 milliards en 2025.

[145] Le projet Stargate des Etats-Unis affiche une ambition d'infrastructure de 500 milliards de dollars. La dépense publique chinoise pour l'IA est estimée à 56 milliards de dollars pour 2025, adossée à des fonds nationaux et des contrôles à l'exportation.

Mistral a massivement recours à la distillation pour proposer des modèles moins coûteux et réduire les coûts en interne, sans que toutefois cette opération ne permette de rattraper le temps sur les concurrents.[146]

Par ailleurs, **l'adaptation des modèles *open-weight* doit être traitée comme un levier de souveraineté à part entière**. En effet, alors que le *hardware* est monopolisé et devient un goulot d'étranglement, la Chine déplace le terrain de la bataille en menant la course de l'intelligence algorithmique. L'exemple de DSpark, le module logiciel publié par DeepSeek, qui accélère l'inférence jusqu'à 85% sans réentraînement et sans puce illustre cette brèche qui neutralise l'avantage matériel des États-Unis, et la dépendance qui en découle.

> *« Si on utilise des modèles* open source *qui sont chinois, ou qui ne sont pas nationaux, ces modèles n'auront pas nos valeurs, et potentiellement il y a des backdoors à l'intérieur. »*
>
> Charbel Raphaël Segerie[147]

Recommandation n°27 (CT) : **Affirmer la spécialisation de l'INESIA sur les risques systémiques**

(a) Investir dans l'IA défensive : orienter une part de l'effort public vers les usages défensifs comme la détection, réponse, remédiation, durcissement ;

(b) Maintenir un suivi continu des capacités (cyber et NRBC) des modèles de frontière, en lien avec la communauté du renseignement, y compris entre les publications majeures ;

(c) Doter l'INESIA et les instituts d'évaluation d'un socle technique partagé pour auditer les modèles *open-weight* et en corriger les distorsions héritées ;

(d) Produire des avis gradués par seuils, à destination des administrations, des développeurs et des secteurs d'infrastructures critiques ; chaque franchissement de seuil déclenchant des recommandations de durcissement de posture ;

(e) Publier annuellement un tableau de bord public de la dépendance cognitive : part des requêtes IA européennes traitées par des modèles sous juridiction étrangère ; part des modèles déployés dont les spécifications comportementales sont publiques ; nombre de chercheurs en alignement travaillant en Europe.

D'autre part, garantir que l'interruption d'un système reste elle-même une option praticable : imposer aux opérateurs d'importance vitale le maintien de procédures de fonctionnement sans IA pour assurer leurs missions essentielles.

La résilience n'est pas l'art de survivre aux désalignements des systèmes d'IA, mais plutôt la capacité à les corriger avant qu'ils ne deviennent des risques systémiques.

[146] Assemblée nationale, Commission d'enquête, Audition d'Arthur Mensch, op.cit., p.9.

[147] Charbel-Raphael Segerie, CeSIA, audition

Recommandation n°28 (CT) : **Institutionnaliser des exercices nationaux et européens de perte de maîtrise,** sur le modèle des exercices de crise cyber

Chaque incident, quasi-incident et exercice alimente en retour les seuils, les standards, les primes d'assurance et la doctrine de reprise de maîtrise.

(a) Maintenir obligatoirement un accompagnement non numérique (guichet, téléphone, courrier) de qualité équivalente (y compris dans les délais de traitement) pour tout service public dématérialisé, avec dispositifs renforcés dans les zones de déserts numériques ou de déficit en littéracie numérique. Les Associations des Maires de France et des Maires ruraux de France, ainsi que France Services et le Défenseur des droits, peuvent être mobilisés pour garantir le droit au recours ;

(b) Instituer des pouvoirs d'urgence gradués de reprise de maîtrise : restreindre ou suspendre temporairement l'accès public à un modèle, y compris par injonction d'arrêt de l'activité de calcul, enjoindre à un fournisseur de modèles de frontière de prendre sans délai des mesures de prévention, d'atténuation ou de réponse pendant la phase aiguë d'un incident ;

(c) Actualiser les plans de continuité pour des dégradations prolongées, et anticiper des ruptures s'étendant sur plusieurs semaines ou mois en définissant les fonctions prioritaires, les modes dégradés, les procédures manuelles et les sites alternatifs ;

(d) Projeter une capacité mutualisée de sécurité vers les opérateurs sous-dotés : déployer des ingénieurs de sécurité auprès des hôpitaux, collectivités territoriales, opérateurs d'eau et d'énergie et PME critiques, mutualiser des SOC régionaux, instituer un centre mutualisé de supervision des agents IA déployés, prestations mutualisées de *red teaming* et de supervision à l'exécution ;

(e) Pré-positionner l'entraide *via* des accords d'assistance mutuelle entre opérateurs pour mutualiser la mobilisation d'équipes de réponse et d'équipements ; et constituer des réserves stratégiques de composants critiques ;

(f) Porter au niveau européen un élargissement substantiel de la réserve de cybersécurité de l'UE (36 M€) aux incidents directement liés à l'IA.

Une perte de maîtrise ne s'arrêtera pas aux frontières, d'autant que la plupart des systèmes concernés sont développés hors Europe.

Recommandation n° 29 (MT) : **Porter la dimension internationale de la résilience**

Proposer la structuration d'une instance internationale de coordination technique, de notification d'urgence des incidents, et de supervision des normes, sur le modèle du *Financial Stability Board* (FSB) ou de l'Agence internationale de l'énergie atomique (AIEA).

À une toute autre échelle, les collectivités territoriales sont les véritables pivots d'une résilience locale, aux manettes de services essentiels de la nation au quotidien, qui reposent aujourd'hui sur la gestion de la donnée : les réseaux d'énergie, l'action sociale, les mobilités, la transition environnementale, etc. À titre d'exemple, l'initiative lancée *via* le programme Territoires d'IA, par la Caisse des Dépôts consistant à déployer 5000 licences d'un outil souverain (Mistral Vibe Compute) pour aider les secrétaires de mairies rurales de la **région Grand Est** va dans le bon sens.

Permettre que la gestion énergétique d'une collectivité ou que les données de mobilité d'une autre collectivité puissent dépendre de fournisseurs privés ou d'entités étrangères représente, en effet, un risque de rupture de service public majeur, voire un risque de rupture politique. À cet égard, la proposition du MEDEF de subventionner l'achat de tokens, *de facto* américains, dans le cadre d'un « crédit d'impôt IA », atteste pour votre rapporteur *a minima* d'un certain manque de vision stratégique.

Certaines pistes méritent d'être sérieusement explorées pour garantir une continuité de la sûreté en mode dégradé et assurer un possible leadership sur le coup d'après. À cet égard, l'informatique moléculaire, et notamment le stockage sur ADN permet de conserver hors ligne certains actifs numériques particulièrement critiques : clés racines, éléments de reprise d'activité, configurations de confiance ou données sensibles de long terme. La société française Biomemory, qui a racheté la société américaine DNA Catalog, développe une solution interfaçable qui nécessite 100 fois moins d'espace que le stockage numérique. En raison de sa vitesse d'encodage et de lecture plus lente, il s'agit d'une couche complémentaire aux infrastructures opérationnelles de cybersécurité, et non d'un substitut aux HSM, KMS ou dispositifs existants.

La résilience puise une grande part de sa force dans la croyance en un récit collectif. L'adhésion et l'embarquement des utilisateurs sont une clé du succès de la reprise en main, afin d'éviter un contournement par le marché ou un verrouillage par les *gatekeepers*. En effet, l'adoption rapide de l'IA, directement en BtoC, a installé des dépendances d'usage : ChatGPT (OpenAI) a franchi le cap du milliard d'utilisateurs mensuels actifs en mai 2026.

Horizon de remembrement stratégique : développer une offre européenne robuste et plurielle de l'IA alignée

Une gouvernance mature ne se contente pas d'éteindre des incendies : l'alignement doit être la boussole de notre politique de maîtrise, dont la filière scientifique, technique et industrielle constitue l'un des principaux instruments, afin d'enrichir et de renforcer notre écosystème.

Cette trajectoire s'appuie sur des garanties démontrées (évaluations, audits, certifications), ce qui la distingue des ambitions d'autonomie stratégique restées déclaratives, ainsi que des dimensions démocratiques et éthiques, qui relèvent davantage de la volonté politique.

Dans cette perspective, la contrainte réglementaire européenne, souvent présentée comme un frein, constitue le moteur d'un marché intérieur de la garantie démontrée, dont les standards, les auditeurs, les certificats et les outils ont vocation à s'exporter vers l'ensemble des marchés régulés.

S'il impose un cadre réglementaire, l'Etat doit aider l'écosystème à s'organiser ensuite dans une logique d'émulation permettant compétition, coopération et innovation. L'écosystème est mobilisé sur l'ensemble de la chaîne pour fournir les composants et maillons essentiels, à l'instar d'une constellation industrielle et dans une logique d'équipementiers.

> **Illustration : EuroStack**
>
> Eurostack est un collectif de dirigeants d'entreprises technologiques européennes opérant dans différents niveaux du stack numérique : du cloud logiciel en passant par la connectivité et l'IA, convaincus que l'Europe doit inverser la tendance qui l'a transformée en une « colonie numérique » dominée par les *hyperscalers*, et retrouver son rôle de fournisseur de services aux clients européens par l'activation de la demande, notamment privée.

La priorité est de faire émerger une demande explicite et solvable pour faire de l'alignement un différenciant économique et un facteur d'adoption de l'IA. À cet égard, le partenariat stratégique pluriannuel conclu par Mistral AI avec Microsoft (*cf supra* p.48), témoigne surtout du cruel manque de débouchés et d'une demande dynamique pour les solutions européennes, même si des mouvements sont en cours au sein de certains groupes industriels européens.

> **Illustration : European Champions Alliance (ECA)**
>
> L'*European Champions Alliance* est une association à but non lucratif qui défend la souveraineté technologique européenne et favorise la coopération numérique au sein de l'écosystème. Une initiative concrète en constitue un exemple pertinent : elle rassemble une vingtaine de groupes européens, dont ArianeGroup, Arte, Atos, Bosch, Cegeka, Deutsche Telekom, Givaudan, IONOS, Rehau, SAP et Siemens, pour d'une part, faciliter le passage à l'échelle de LLM européens fondés sur l'acquis communautaire et, d'autre part, formaliser une « souveraineté de l'inférence » grâce à l'adaptation locale des modèles. Ces groupes se fédèrent pour définir leur demande, partager leurs cas d'usages et accélérer une disponibilité d'offre.

La bonne approche est donc progressive, itérative, et réversible : il ne s'agit pas de construire une forteresse, mais plutôt de **faire vivre notre forêt** [148]***via* un narratif fort et mobilisateur** (cf. Scénario idéal, annexe), **afin d'enraciner la confiance par les perspectives de débouchés et la réalité des commandes (publique et privée**).

> Recommandation n°30 (MT) : **Consolider une Base Industrielle et Technologique du Numérique et de l'IA (BITN-IA) : la « forêt intelligente »**
>
> Conçue comme une architecture agrégeant les moyens publics, industriels, scientifiques et financiers nécessaires dans une logique écosystémique, cette Base permettrait d'articuler l'écosystème entre innovation ouverte (foisonnement, agilité, niches compétitives) et industrialisation dirigée (convergence, masse critique) *via* des outils de financement :

[148] La métaphore de la forêt ne doit pas suggérer que l'IA obéirait à des lois naturelles indépendantes de choix humains. Les systèmes d'IA sont conçus, financés, entraînés, déployés et gouvernés par des acteurs humains identifiables. La forêt est une image d'interdépendance et d'adaptation, non une excuse à la fatalité.

(a) Lancer des fonds d'infrastructure et des préachats (*offtakes*) pour les infrastructures coûteuses, notamment les racks de semi-conducteurs ;

(b) Faciliter les passerelles entre *deeptechs* civiles et programmes de défense (*continuum*) : le marché de la défense peut offrir des débouchés structurants mais impliquer un verrouillage pour des questions doctrinales (ex : les SALA[149], exclus du RIA), d'acceptabilité et de financement croisé ;

(c) Instaurer un dispositif d'avances remboursables non dilutif : CDC/Bpifrance pour couvrir jusqu'à 50% des besoins de trésorerie ;

(d) Créer un volet « IA de confiance, socle et alignement » dans les dispositifs de financement (Bpifrance, initiative Tibi, FEI), incluant le capital-croissance, afin d'ancrer en Europe les futurs leaders de l'évaluation, de l'interprétabilité et de l'audit de modèles, et d'éviter leur migration/prédation au moment de leur passage à l'échelle. Cette meilleure articulation entre amorçage et industrialisation s'effectuera à l'aune d'ETCI 2.0 ;

(e) Co-construire avec France Invest, France Deeptech et Invest Europe une grille de *due diligences* « alignement » légère et standardisée, pour diffuser l'exigence dans toute la chaîne de financement sans créer de charge documentaire disproportionnée pour les entreprises ;

(f) Reventiler certains crédits d'impôt (CII/CIR) pour stimuler un amorçage incitatif[150], *via* de potentiels amendements de votre rapporteur en ce sens au projet de loi de finances pour 2027. Un mécanisme correcteur de l'effet régressif de la mise en conformité du RIA (*cf. supra*) pourrait être envisagé lorsque ce coût dépasse 2% du chiffre d'affaires annuel de l'entreprise ;

(g) Réorienter une partie de l'épargne privée, notamment vers des placements de plus long terme et/ou par la simplification des outils de capital-risque.

Cette BITN-IA n'a de sens que si elle est soutenue par les acteurs économiques européens, qui pourraient initier les premières adoptions de solutions d'alignement technologiquement satisfaisantes.

Le 10 juillet 2026, la Banque européenne d'investissement (BEI) a confirmé la participation de plusieurs investisseurs institutionnels au deuxième millésime de l'initiative *European Tech Champion* (ETCI), afin de créer un **effet de levier par le financement privé, lever 80 milliards d'euros, et boucher le « *scale-up gap* ».** La plateforme doit irriguer plus de 1 500 entreprises européennes en expansion, via plus de 100 fonds, dont jusqu'à 45 méga-fonds capables d'injecter en moyenne 200 millions d'euros par entreprise. Un autre exemple, côté privé, est le programme européen pour la souveraineté, annoncé par Generali le 20 juillet 2026, et structuré comme un « **fonds de fonds » doté de 300 millions d'euros**, destiné aux PME et infrastructures technologiques jugées stratégique à l'échelle européenne.

[149] Système d'arme létal autonome

[150] AFNOR Normalisation, « Stratégie Française de normalisation 2025-2030 : la normalisation, levier d'innovation, de souveraineté et d'influence pour la France », édito de Sébastien Martin, Ministre délégué chargé de l'Industrie, p.3

Par ailleurs, la France doit donner l'impulsion, au niveau européen, **pour que la commande publique soit un levier d'amplification de la dynamique privée et un instrument de structuration d'un marché solvable pour les solutions alignées.**

Recommandation n° 31 (MT) : **Faire de la commande privée et publique un levier structurant pour la filière française et européenne de l'IA alignée**

(a) Élaborer un guide de clauses-types avec des critères qualitatifs de l'alignement, modulés sur les articles 9 à 15 du RIA (système de gestion des risques, gouvernance des données, documentation technique, tenue de registres, transparence, contrôle humain, et robustesse des SIA à haut risque), et un volet environnemental[151] (durée de vie des équipements, intensité de calcul, sobriété énergétique) pour flécher vers des solutions conformes.

(b) Une bonification de notation serait accordée aux solutions documentant leur empreinte sur le référentiel général d'écoconception de services numériques (RGESN), sur le référentiel général pour une IA frugale (*cf. infra*) ou encore sur certains labels et certifications favorisant le réemploi ou l'acquisition de matériels reconditionnés, et celles cartographiant leurs vulnérabilités *via* l'Indice de Résilience Numérique (IRN). À l'échelle européenne, le "*Score Sovereignty Framework*", barème conçu pour mesurer la souveraineté des prestataires de cloud, peut être étendu aux systèmes d'IA afin d'orienter les achats publics de l'administration européenne ;

(c) Mobiliser les dispositifs existants : l'achat innovant[152], qui permet des marchés jusqu'à 100 000 euros sans publicité ni mise en concurrence-, et l'allotissement renforcé dans les marchés IA supérieurs aux seuils européens, avec des lots dimensionnés pour permettre la candidature de startups et PME ;

(d) Orienter de façon préférentielle la recherche partenariale en IA vers les entreprises françaises, notamment *via* le partenariat d'innovation[153] ;

(e) Promouvoir la préférence européenne au Conseil de l'UE, dans le cadre des discussions sur le *Cloud and AI Development Act* ;

(f) Exiger que tout projet d'infrastructure d'IA financé, soutenu, et déployé par la puissance publique (data centers, capacités de calcul) intègre une clause de substitution / réversibilité par du matériel européen, dès qu'une solution souveraine équivalente existe.

Le vecteur juridique envisageable est une circulaire du Premier ministre, sur le modèle « Cloud au centre » du 5 juillet 2021, complétée par une doctrine d'achat IA portée par la DAE et la Dinum, et un guide pratique des clauses contractuelles types incluant le volet environnemental.

[151] Directive 2014/24/UE du Parlement européen et du Conseil du 26 février 2014 sur la passation des marchés publics, articles 67 à 69 relatifs aux critères d'attribution incluant les caractéristiques environnementales ; transposition en droit français par le Code de la commande publique.

[152] Code de la commande publique, article R. 2122-9-1.

[153] Code de la commande publique, article R. 2124-3 : co-développement avec des PME sur trois à sept ans.

Le Conseil national de la commande publique (CNCP), créé en février 2026, pourra investir ce levier, dans un cadre institutionnel clarifié par la création de l'Ariane, et à la lumière des conclusions de la commission d'enquête sénatoriale, initiée par le groupe « Les Indépendants-République et Territoire » (LIRT) sur les coûts et les modalités de la commande publique.

Cette mise en œuvre suppose toutefois de changer les mentalités et de renforcer les compétences des acheteurs publics et des chefs de projet, afin de limiter l'asymétrie d'information avec les fournisseurs et d'éviter que les administrations ne soient orientées vers des solutions plus puissantes, plus coûteuses et plus consommatrices que nécessaire. De plus, les comités de direction et les conseils d'administration doivent disposer d'un cadre préalable pour déterminer, avant l'engagement, ce qui demeure sous responsabilité humaine directe, les dépendances qu'ils acceptent et les conditions de leur réversibilité.

En parallèle, il convient de **stimuler l'émergence d'une demande explicite et solvable**, afin de faire de l'alignement un différenciant économique et un facteur d'adoption, notamment en matière de gestion du risque et de dialogue social.

Recommandation n°32 (CT) : **Objectiver des critères d'alignement observables et exigibles par les acheteurs publics**

(a) Formaliser des notations publiques et comparables de la maturité des pratiques de gestion des risques, établies sur audits indépendants (à l'image de la notation financière) ;

(b) Publier des politiques de sécurité de frontière (*frontier safety frameworks*) des développeurs opérant sur le marché européen : seuils de non-délégation, mesures et engagements rendus opposables et vérifiables ;

(c) Prévoir dans tout marché public, portant sur un SIA, une obligation contractuelle de portabilité des données et des paramétrages vers un système alternatif, dans un format documenté et exploitable sans dépendance technique excessive au fournisseur sortant ;

(d) Pousser, lors du prochain G7, en faveur de la création d'un organisme intergouvernemental qui recenserait les fournisseurs de modèles à haut risque non coopératifs dans une liste noire, avec des contremesures recommandées ; sur le modèle du Groupe d'Action Financière (GAFI) pour le blanchiment d'argent (sorte de *name and shame* avec gel des avoirs) ;

(e) À défaut d'accord du fournisseur sur les clauses d'audit, de portabilité, ou de réversibilité, privilégier les modèles *open-weight* auditables, à performance comparable ou légèrement dégradée, dont l'adoption croissante dans les administrations et les entreprises crée un besoin de vérification spécifique.

Dans un contexte d'une demande, dans le cloud et l'IA, historiquement monopolisée par les hyperscalers américains, la part de marché des acteurs européens a plutôt tendance à se compresser. Alors que pour tout milliard d'euros que génère OVH, Amazon Web Services (AWS) en engendre 18 en Europe, le seul challenger européen sur les accélérateurs IA, Graphcore, a été racheté par SoftBank en 2024[154].

[154] Christophe Picou, Contribution écrite : » Gouverner la dépendance : Diagnostic et cadre d'action pour les dirigeants européens face à l'IA », 2026, p.4.

Ces chiffres ne décrivent pas un retard conjoncturel, mais une réalité structurelle inscrite dans les compétences et les cadres juridiques. Selon OVH, si 15% du marché public allait vers des solutions européennes, les capacités d'investissement seraient doublées[155].

Des corrections de trajectoire sont encore possibles, comme l'a illustré l'hébergement du Health Data Hub, de Microsoft Azure vers Scaleway. Un mouvement, encore relativement sous les radars, s'opère au sein d'acteurs privés basculant vers des solutions souveraines, à l'instar du récent choix d'Airbus qui a préféré à AWS cloud Scaleway pour héberger ses données critiques. **L'action publique doit encourager, voire accompagner ce mouvement, qui prolonge la logique des « accélérateurs de demande » introduite par le Chips Act 2.0. La commande publique innovante est le point d'entrée immédiat, mais son plein effet de relais suppose une demande publique structurée dans la durée.**

À cet égard, si la fragmentation du marché est un véritable frein pour le passage à l'échelle, les entreprises européennes sont subséquemment davantage ouvertes aux démarches partenariales.

Le manque d'investissement privé européen bloque l'émergence d'une offre européenne. Dans ce contexte, la Commission européenne a annoncé un train de mesures sur la souveraineté technologique européenne le 3 juin 2026, dont une future capacité européenne d'investissement en fonds propres dans les technologies. **Il paraît primordial que la France s'investisse dans la consultation qui sera lancée avec les États membres et la Banque européenne d'Investissement (BEI).**

Par ailleurs, la Commission européenne devrait présenter des mesures pour **alléger les règles prudentielles des banques**[156]. En effet, la mise en œuvre européenne des accords de Bâle IV relève d'environ 15 % les exigences de fonds propres des grandes banques, quand les États-Unis les abaissent d'au moins 5 %, creusant un écart de compétitivité, et un manque à gagner estimé à 2000 milliards d'euros de capacité de financement chez les principales banques européennes.

Recommandation n°33 (MT) : **Créer un écosystème intégré sur l'ensemble de la chaîne de valeur pour vendre, sous une bannière unifiée, notre « IA alignée »**

(a) Proposer des offres communes (bundles) européennes compétitives : capacités de calcul (ASML, SiPearl, Vsora, Bull…), cloud (Scaleway, OVH, Orange, CleverCloud…), énergie décarbonée (EDF, Schneider…), modèles (Mistral, Black Forest Labs…) ;

(b) Intégrer des objectifs commerciaux dans les missions des ambassadeurs et des conseillers du commerce extérieur + recruter des experts techniques et commerciaux dans les ambassades pour présenter un message unifié de l'offre stack européenne, sur le modèle EuroStack ;

(c) Faciliter la constitution de consortiums industriels et soutenir les candidatures dans le cadre de l'appel d'offres « AI Gigafactory » de la Commission européenne.

[155] Clara Chappaz, ambassadrice pour l'IA et le numérique, audition

[156] Commission européenne, Communique de presse « La Commission présente des mesures visant à renforcer le secteur bancaire européen et à soutenir la croissance », 17 juillet 2026.

Illustration : Consortiums industriels

Des consortiums industriels se sont structurés pour assembler des chaînes industrielles complètes et intégrées :

- ÆTHER, qui réunit des acteurs européens du calcul (2CRSi, Axelera AI), des puces (SiPearl), du cloud (Outscale, Viridien), de l'énergie (Electricité de Stasbourg, Socom, Haffner), et de la construction (Nhood, Equans, Projex), à Strasbourg, pourrait être mis en service dès 2027 ;
- AION met en avant une combinaison de la puissance financière (Ardian), de l'énergie bas carbone (EDF), du cloud souverain (Scaleway), des télécoms (Orange, Iliad), du supercalcul (Bull), du conseil (CapGemini, Artefact), et du déploiement industriel, avec un écosystème élargi de partenaires : Hugging Face, Kyutai, Inria, GENCI, Quandela, Nokia, Vsora, SiPearl, Schneider Electric…
- Eurocommons est une agrégation d'offres industrielles (*top down*), alternative à Atlassian : 15 grands comptes français et allemands développent une solution *open-source* dans le cadre du programme Horizon Numérique 2030 de la Caisse des dépôts. Ce catalogue rassemble des offres interopérables pour identifier des besoins communs et préparer des plans de migration en partageant le coût. Ces migrations doivent être **drivées par la demande** — les besoins exprimés par les DSI, éclairés par une cartographie des dépendances numériques critiques via l'IRN notamment.

Ce foisonnement d'initiatives est à saluer, sous réserve de veiller à ce que la multiplication des agrégations ne s'apparente pas à un nouvel empilement décentralisé de couches.

Pour éviter cet écueil, **la France et l'Europe doivent privilégier une approche intégrée, sur le modèle de la Corée du Sud,** qui a lancé une initiative ambitieuse (environ 320 millions d'euros) de mise en concurrence de 5 entreprises (LG, SK Telecom, Naver, NC AI, et Upstage) au sein d'un consortium, avec des phases éliminatoires, afin de développer des modèles d'IA souverains d'ici 2027. En parallèle, la Corée du Sud a dévoilé un plan colossal, équivalent à plus de 1000 milliards d'euros sur 10 ans pour construire des usines de semi-conducteurs avancés et des data centers. La position dominante de la Corée du Sud dans les semi-conducteurs[157] - fruit d'une politique publique de long terme - lui confère des marges de manœuvre dans les partenariats internationaux (avec AWS, Microsoft, Nvidia).

À cet égard, **l'Union Européenne et la France disposent d'un levier énergétique (puissance électrique décarbonée et abondante) et concurrentiel (marché intérieur**), qu'il convient de mobiliser, notamment pour inciter des partenariats de recherche en « boîte blanche ». **La construction d'une expertise indépendante et d'un marché y afférent appelle des compromis à prioriser**.

[157] Samsung et SK Hynix produisent 90% de la mémoire HBM mondiale.

Recommandation n°34 (CT) : **Pondérer l'accès à différentes ressources et capacités de calcul**

(a) Exercer un droit de regard sur les modèles entrainés sur le sol français (et donc bénéficiant de l'empreinte foncière, du raccordement au réseau…) ;

(b) Réserver une partie des capacités (calcul, énergie) aux acteurs européens, afin d'éviter un effet d'éviction pour d'autres industries électro-intensives : les PME/ETI développant en bacs à sable réglementaires (*cf. supra*) pourraient disposer de procédures d'accès simplifiées via une convention avec le GENCI ;

À cet égard, le « droit de préemption »[158] par le Gouvernement de fréquences radioélectriques nécessaires à l'accomplissement de missions de service public pourrait servir d'inspiration pour instaurer un droit de priorité sur la location des centres de données et l'utilisation de l'électricité décarbonée ;

NB : la compétence d'allocation d'un terrain pour la construction d'un centre de données revient à l'intercommunalité, qui n'est pas forcément consciente des enjeux supranationaux.

(c) Ouvrir l'accès à certaines données spécifiques de marchés régulés contre l'audit des techniques d'alignement proposées par les concepteurs de modèles ;

(d) Soumettre le déploiement d'un modèle des plus avancés à une AMM (autorisation de mise sur le marché), délivrée par l'INESIA (ou l'AI Office), évaluant notamment les externalités négatives, les pratiques anticoncurrentielles éventuelles du concepteur (*hyperscaler)*, et sa conformité aux cadres de protection des données, de transparence algorithmique et d'éthique numérique.

Le nouveau pétrole, c'est la capacité à produire de l'énergie propre à l'échelle industrielle.

> *« La France dispose de 90 térawattheures d'électricité par an. On les utilise, entre autres, pour l'intelligence artificielle. Certains acteurs ont largement les moyens financiers nécessaires pour transformer des électrons en IA, quelle que soit la demande. Ils vont le faire dans les deux prochaines années. Ensuite, il n'y aura plus d'électricité et il faudra créer de nouvelles centrales. On assiste à un phénomène, qui s'accélère à une vitesse folle, de monopolisation de la ressource énergétique européenne. On n'a pas conscience de ce processus pourtant irrémédiable. »*
>
> *Arthur Mensch*[159]

[158] Loi n ° 86-1067 du 30 septembre 1986 relative à la liberté de communication, dite « Loi Léotard », article 26, Modifié par LOI n°2021-1382 du 25 octobre 2021 - art. 33

[159] Assemblée nationale, Commission d'enquête, mai 2026, op.cit., p.8.

La souveraineté n'est pas l'autarcie, mais la naïveté devient obsolète : il faut trouver les moyens de créer un « protectionnisme libéral »[160]. Un gouvernance trop molle peut favoriser des espèces invasives, au détriment de jeunes pousses, qui peinent à émerger, faute des nutriments nécessaires à leur bonne maturité.

La conditionnalité de l'accès aux ressources et aux capacités de calcul attesterait d'un réveil stratégique afin, en conscience, de ne pas laisser se déplacer subrepticement le centre de gravité de nos dépendances et la captation de la valeur économique.

Une fois la gouvernance du calcul consolidée dans une chaîne d'approvisionnement contrôlée par des points européens, la **deuxième étape est d'assurer que la commande (publique et privée) soit réinvestie dans de la R&D sur le territoire européen**, et qu'elle permette l'émergence d'une plus grande diversité de fournisseurs d'IA européens. La présence de Mistral, seule licorne de l'IA européenne[161], nous oblige à insuffler la diversification en ce sens.

Dans une forêt, un arbre immense peut stabiliser un écosystème ou bloquer toute diversité : en ce sens, la gouvernance doit éviter qu'un seul arbre ne monopolise la canopée et les ressources. Nous pourrions nous inspirer de l'Allemagne dont le « High Tech Agenda Germany » (HTAG) établit un objectif de trois modèles d'IA souverains dans des domaines spécifiques et se distinguant par leur sécurité et durabilité d'ici 2029.[162]

À cet égard, **l'agence ARPA dédiée à la sécurité de l'IA par conception** (*cf. recommandation n° 2, supra*) serait un modèle d'économie mixte dont les résultats alimenteraient les méthodes d'évaluation de l'INESIA, la normalisation (CEN-CENELEC JTC21) et l'écosystème industriel (BITN-IA). De plus, elle donnerait corps à la définition franco-allemande de la souveraineté numérique. Les annonces du 26e Conseil des ministres franco-allemand, le 17 juillet 2026, sur l'IA de frontière, confortent votre rapporteur en ce sens, en traçant un pas de plus vers la coopération renforcée pour réduire, ensemble, nos dépendances grâce au renforcement de l'écosystème européen.

> *« L'évaluation permettra d'innover et de se remettre dans la course à la création. On a les compétences, en France et en Europe, pour gagner cette course de l'évaluation*. »
>
> Jean-Michel Loubes[163]

[160] Yann Lechelle, audition du 2 juin 2026

[161] Future of Life Institute, AI Safety Index, Summer 2026 Edition : si Mistral obtient la pire note des 9 fournisseurs de modèles d'IA avancés évalués sur leurs pratiques de sécurité (F), il s'agit du seul fournisseur ni américain ni chinois. De plus, son entrée dans le classement témoigne de son évolution fulgurante.

[162] Etude comparative internationale, Service pour la science et la technologie de l'Ambassade de France en Allemagne, p.2

[163] Jean-Michel Loubes, Audition du 10 juin 2026

Chapitre 3 - L'alignement comme stratégie ricardienne de différenciation stratégique

La solution : ne surtout pas chercher à copier ce qui est déjà fait, mais construire une alternative et préparer le coup d'après.

L'enjeu majeur est d'identifier les segments et briques maitrisables, ainsi que les voies par lesquelles nous pouvons diffuser les standards, sur l'ensemble des tableaux : public/privé, civil/militaire, national/européen/international.

Le débat public confond souvent deux réalités économiques très différentes. D'un côté, l'infrastructure (data centers, parcs de calcul), exposée à l'obsolescence rapide des puces. De l'autre, **la technologie (conception de semi-conducteurs, modèles, briques d'alignement et d'interprétabilité) : c'est cette dernière, bien plus rare et bien plus créatrice de valeur, qui est réellement différenciante**. Subventionner massivement des data centers ne crée pas de souveraineté : cela finance un actif amortissable et souvent largement importé. La priorité, publique comme privée, devrait porter sur la couche rare — la technologie — et non sur la couche banalisable — l'infrastructure.

L'intelligence industrielle : le dernier rempart

L'IA dynamite l'édifice logiciel sur lequel s'est construite la puissance américaine en plaçant l'infrastructure au cœur de la compétition.

La différenciation européenne pourrait donc passer par **le dépassement du paradigme des LLM** centralisés et le développement de niches stratégiques d'IA appliquée.

Face à l'avantage des grands acteurs extra-européens dans le segment du développement des modèles de fondation généralistes et de frontière, la France et l'Europe ont l'opportunité de se concentrer sur les **segments où la maîtrise de l'alignement constitue un différenciateur économique direct : la certification industrielle, le déploiement embarqué, la vérification, la frugalité, ainsi que l'alignement des systèmes autonomes physiques où la dimension sectorielle est dominante.**

Recommandation n°35 (MT) : **Structurer l'écosystème européen de spécification-vérification**

La vérification est un angle mort des stratégies du réseau des AI Safety Institutes. L'Europe détient une force historique dispersée, que l'AI Office peut coordonner : Inria (CompCert, Frama-C), Chalmers (CakeML), TU Graz, MPI-SWS, IMDEA, ETH Zürich, premier laboratoire industriel commun Inria–Mitsubishi (FRAIME) ;

(a) Placer les mécanismes de vérification ancrés dans le matériel (eg. mécanismes *on-chip*) comme racine de confiance physique complétant les audits logiciels ;

(b) Tirer profit de la position européenne dans la chaîne d'approvisionnement (ASML, etc.) comme levier de négociation diplomatique pour inscrire des exigences de vérification dans les accords internationaux.

Dans les domaines où le savoir-faire considéré présente un caractère de double usage, notamment en biosécurité[164], il n'est pas souhaitable de se reporter seulement aux frameworks spécifiques[165] mis au point par les principaux concepteurs d'IA.

L'évaluation du risque NRBC-e est une catégorie distincte qui doit être conduite par des experts en sciences fondamentales des domaines concernés. Un cas d'usage particulièrement stratégique pour la France est l'IA appliquée à la conception et à l'exploitation des réacteurs nucléaires : optimisation de design (codes neutroniques, géométrie de cœur, gestion de combustible), assistance à l'exploitation (analyse de signaux, maintenance prédictive), simulation accélérée par des modèles de substitution pour des calculs coûteux.

Recommandation n°36 (CT) : **Mise en place de pôles d'expertise sectoriels de l'IA appliquée au sein de l'INESIA**

Premiers marchés naturels : énergie et infrastructures critiques, robotique industrielle et collaborative, santé, défense.

(a) Constituer une contre-expertise indépendante systématique sur les risques qu'une IA confère à un acteur les capacités d'actions dans les domaines NRBC-e ou permette les usages duaux ;

(b) Associer les autorités de sûreté (ASNR, AIEA) car les enjeux dépassent la compétence d'un acteur industriel isolé ;

Chaque pôle élaborerait méthodologie, benchmarks et service de certification avec une vue d'ensemble sur les sous-traitants : quantification des erreurs (prédiction conforme, inférence bayésienne), intégration de contraintes physiques dans l'architecture ou l'objectif, recalibration post-hoc, couplée à des ontologies pour les agents scientifiques (projet Exa-DI du PEPR NumPEx), analyse des synthèses d'acides nucléiques.

Les conséquences d'un désalignement sont exacerbées, et même irréversibles, dans les systèmes d'IA autonomes physiques : modèle vision-langage-action (VLA) embarquées dans des robots ou véhicules ou dans des applications critiques.

Un système peut ainsi paraître performant sur la base des métriques publiques, et cesser de l'être après portage en embarqué, **précisément le segment où l'industrie européenne peut disposer d'un avantage compétitif (efficacité énergétique, indépendance vis-à-vis du *cloud* extra-européen).** Le verrou industriel européen sur l'IA embarquée se trouve ainsi directement lié à la maîtrise de la robustesse de l'alignement aux transformations de déploiement.

[164] Gotting et al., 2025 : des modèles d'IA génèrent déjà des protocoles de virologie supérieurs à ceux de 94% des experts humains.

[165] Frameworks visant la Preparedness : désigne la capacité organisationnelle et technique à anticiper, évaluer, contenir et gouverner les risques sévères associés aux modèles avancés.

<u>Illustration :</u> IA embarquée : confidentialité et frugalité

Les capteurs de vision neuromorphique évènementielle captent uniquement des variations de lumière : 10 000 images/seconde. Cette forte réactivité permet une frugalité énergétique avec la détection passive, robustesse en conditions réelles et lumineuses variées (brouillage), et protection de la vie privée par conception, car il n'y a pas d'image directement exploitable.

Prophesee, société française créée en 2014 et issue des travaux de l'Institut de la Vision (Sorbonne Université, CNRS, Inserm), est leader mondial dans ce domaine, étant la seule entreprise au monde à avoir industrialisé des capteurs de vision évènementielle à grande échelle, avec 10 ans d'avance sur ses concurrents tout juste émergents.

Confidentialité, frugalité, fiabilité sont tout autant de rimes avec acceptabilité ; atouts non négligeables quand la confiance est devenue la denrée la plus rare de notre siècle.

Le chantier de la standardisation de l'IA embarquée est encore largement ouvert[166], et sa maîtrise constitue un terrain naturel, à la lumière de la culture française (et européenne) de sûreté des systèmes critiques, mais également à l'aune de l'objectif de développer une filière européenne de l'IA physique inscrite dans le CAIDA[167].

<u>Recommandation n° 37 (MT) :</u> **Adapter les cadres de régulation et de certifications sectoriels aux composants apprenants**

(a) Porter, au sein du Conseil de l'UE, l'inscription explicite des technologies de perception et d'imagerie parmi les technologies critiques soutenues dans le cadre de la proposition de règlement européen dite « Chips Act 2.0 » du 3 juin 2026 ;

(b) Intégrer le matériel et les capteurs, ainsi que la frugalité énergétique et la protection de la vie privée *by-design* comme critères constitutifs de l'alignement dans le cadre des travaux de normalisation ;

(c) Inclure les composants apprenants dans les applications critiques où les certifications sectorielles existantes sont à étendre : le nucléaire (IEC 61513), l'aéronautique (DO-178C), l'automobile (ISO 26262), le médical (IEC 62304), et la sûreté de fonctionnement industrielle (IEC 61508) ;

La consolidation d'un acteur tel que Prophesee, acteur issu de la recherche publique et qui maitrise sa technologie de bout en bout, participe directement de cette capacité d'influence normative.

[166] SaferAI : au sein du ISO SC42, le sujet de l'IA embarquée est abordé principalement via des contributions d'experts, sans qu'il n'y ait véritablement de proposition de standardisation pour l'instant, mise à part une proposition de standardisation de la qualité des évaluations des "embodied AI systems", qui n'a pas abouti pour l'instant.

[167] Projet de règlement européen *Cloud and AI Development Act*, présenté le 3 juin 2026.

La frugalité : la nouvelle valeur ajoutée des SIA

La course mondiale à l'IA est aujourd'hui dominée par les très grands modèles, dont l'entraînement requiert des centaines de millions de dollars et une consommation énergétique élevée. **En valorisant ses atouts (recherche en mathématiques appliquées, électricité décarbonée, vivier de talents, présence industrielle dans les secteurs régaliens et sectoriels), la France peut se spécialiser sur l'IA frugale, sobre et de contexte ; également davantage adaptée aux besoins des marchés émergents qui constituent sa zone d'influence**.

L'analyse de cycle de vie (ACV) est le cadre de référence[168] permettant d'évaluer l'empreinte des SIA, en prenant en compte l'ensemble du cycle de vie du service, depuis la fabrication des terminaux, serveurs et équipements réseau jusqu'à leur usage, leur hébergement, leur maintenance et leur fin de vie. Dans le contexte français, l'évaluation doit, en outre, tenir compte des conditions concrètes d'hébergement, notamment du recours au cloud, de la localisation des traitements, du mix électrique, de la mutualisation des infrastructures et de la disponibilité de données fiables fournies par les prestataires.

Mistral AI est la première entreprise d'intelligence artificielle qui ait procédé à une analyse du cycle de vie de ses modèles, en 2025, avec l'Ademe, l'Agence de la transition écologique, et Carbone 4.

Cette analyse invite à raisonner à partir du service effectivement rendu, afin de comparer différentes solutions techniques, y compris des alternatives moins automatisées ou moins intensives en ressources, en raisonnant sur l'ensemble des dimensions d'un projet, plutôt que d'envisager *a posteriori* des compensations.

Plus largement, la sobriété implique aussi de limiter les requêtes inutiles, de désactiver les fonctionnalités accessoires et d'encadrer les usages exploratoires lorsqu'ils deviennent massifs.

> Une étude conjointe de l'UNESCO et de l'University College London, présentée au sommet AI for Good, à Genève, indique que des ajustements de conception des LLM (modèles plus petits, interactions plus courtes, compression) peuvent réduire jusqu'à 90 % leur consommation énergétique sans perte de performance.

Investir dans une IA plus durable et éthique pour limiter l'empreinte carbone des services basés sur ces modèles est un avantage comparatif futur, mais également un facteur de souveraineté et d'autonomie, en ce qu'il réduit la dépendance à des infrastructures surdimensionnées.

Si des initiatives isolées sont mises en place pour évaluer l'impact environnemental des SIA proposés par les prestataires (ex : la Région Bretagne[169]), une structuration systémique et harmonisée de cette évaluation est nécessaire afin d'en faire un avantage comparatif, exportable à l'international.

[168] International Organization for Standardization. ISO 14040:2006 – Environmental management – Life cycle assessment – Principles and framework. 2006; International Organization for Standardization. ISO 14044:2006 – Environmental management – Life cycle assessment – Requirements and guidelines. ISO, 2006.

[169] Région Bretagne. Plan de transition bas carbone 2026–2028. Conseil régional de Bretagne, 2 avr. 2026.

Plusieurs outils permettent déjà de mesurer ou d'estimer une partie de l'empreinte environnementale des services numériques publics, sans qu'aucun ne couvre à lui seul l'ensemble du cycle de vie : BoaviztAPI, CodeCarbon, EcoLogits, Green Algorithms ou MLflow peuvent documenter certains aspects des équipements, des infrastructures, des calculs, de l'entraînement ou de l'exécution des modèles d'intelligence artificielle.

L'enjeu est donc d'articuler ces outils dans un cadre méthodologique commun et de les relier aux décisions concrètes d'achat, de conception et de déploiement (cf. commande publique *supra*).

Recommandation n°38 (CT) : **Faire de la sobriété numérique et de l'IA frugale l'axe différenciant de l'alignement** : *PM, ADEME, Ariane, DITP, DAJ, Commissariat général au développement durable*

(a) Inscrire le principe d'alignement soutenable comme critère obligatoire dans les cahiers des charges des appels d'offres des marchés publics ;
(b) Intégrer une évaluation environnementale graduée (simplifiée pour les projets courants, analyse de cycle de vie complète au-delà de seuils de significativité) dans toute évaluation d'impact éthique et d'incidence algorithmique, préalable à un déploiement ;
(c) Rendre obligatoires la mesure et la publication annuelle de l'empreinte carbone des systèmes numériques publics, différenciée par phase de cycle de vie, selon une méthodologie commune ADEME-Dinum ;

(d) Instituer une mission permanente de suivi de l'empreinte environnementale du numérique public chargée de consolider les données et de diffuser des clauses types.

La France pourrait agir efficacement dans le cadre tracé par le RIA en assumant une spécialisation sur l'IA frugale, sobre, sectorielle et exportable. Cette pluralité des approches est la clé pour garantir la diversité de l'écosystème, et donc plus largement, de la biodiversité.

Enfin, les sociétés européennes empruntent la voie d'architectures alternatives, en dépassant le paradigme des LLM et de l'IA générative.

Illustration : les *world models* : la remise en cause des LLM

Les *world models* sont définis comme des SIA capables de simuler des représentations du monde physique, afin de prédire les conséquences des actions ; contrairement aux LLMs et leur propension à générer des hallucinations. La startup AMI Labs fondée par Yann Le Cun se spécialise dans cette méthodologie inédite, principalement basée sur la recherche fondamentale, à contre-courant des optimisations menées par les autres grands fournisseurs. Les *world models* peuvent contribuer à atteindre des architectures plus contrôlables, plus ancrées, et plus évaluables.

Cette approche est particulièrement pertinente pour l'IA physique : robotique, systèmes industriels, énergie, santé, mobilité, logistique… domaines qui ne peuvent se limiter à des sorties textuelles indésirables.

Illustration : la remise en cause du paradigme du *tokenizer*[170]

L'IA, parce qu'elle génère du contenu langagier, façonne les représentations culturelles, la langue, l'éthique ; et peut en ce sens amplifier les écarts et la dualité des valeurs[171]. Cette médiation de l'information est une vulnérabilité puisque, faute de solution alternative européenne, nous sommes contraints d'adopter les modèles extra-européens et d'hériter des biais qui en découlent. C'est donc un enjeu de diversité linguistique, de justice cognitive, et de souveraineté technologique.

Certains projets, portés par Mistral ou Bloom, construisent des tokenizers multilingues, afin de rééquilibrer la représentation des corpus. À cet égard, les travaux conduits au Québec, sur l'articulation de la sécurité de l'IA avec les programmes d'inclusion linguistique, sont à suivre avec intérêt.[172] L'architecture de LLM sans *tokenizer* développée par Aleph Alpha, société allemande, est en ce sens prometteuse pour mieux gérer les spécificités linguistiques et sectorielles des modèles, avec une réduction de 70% de l'empreinte carbone de la phase d'entraînement.

Illustration : l'informatique moléculaire

Le stockage sur ADN (*cf. supra,* p.73) en constitue aujourd'hui l'une des premières applications, mais le domaine ouvre plus largement des perspectives de stockage, de recherche et, à terme, de traitement de l'information reposant sur des architectures différentes du silicium. À ce titre, l'informatique moléculaire semble illustrer le type de pari technologique qu'une ARPA (*cf. supra*, p.18) pourrait soutenir : forte interdisciplinarité, verrous scientifiques et industriels importants, horizons de maturation longs, potentiel de rupture en matière de densité, de sobriété, de résilience et de souveraineté technologique.

Ce changement de paradigme implique la nécessité d'adopter une **approche hybride de formation de nos ingénieurs** pour combiner leurs connaissances en IA (modèles, data, MLOps) avec les enjeux de sûreté/sécurité, de réglementation et d'éthique, et de compréhension des architectures industrielles pour éviter tout verrouillage (*lock-in*). Au-delà, les grands groupes industriels sont unanimes dans leur soutien à la perspective d'un **diplôme d'ingénieur européen *by design***[173], afin de rendre les formations plus lisibles et attractives. Ce projet, que le rapporteur soutient depuis ses prémices dans le cadre de l'initiative d'Université de technologie européenne (EUT+) est un impératif pour renforcer la souveraineté technologique et améliorer la compétitivité industrielle.

[170] Processus de découpage du texte en unités élémentaires et statistiques (tokens) que le LLM peut traiter.

[171] Audition de Nina Landes, coordinatrice nationale pour l'IA, 3 juin 2026

[172] Etude comparative internationa, Canada : Programme des langues autochtones de Patrimoine canadien, le Programme des technologies linguistiques autochtones du Conseil national de recherches et l'initiative FLAIR de Mila pour les communautés autochtones.

[173] Suite au colloque organisé par le rapporteur, Vanina Paoli-Gagin, au Palais du Luxembourg, le 29 juin 2025, Olivier Ginez, Directeur général de l'enseignement supérieur et de l'insertion professionnel (DGESIP) a annoncé le lancement d'expérimentations dans les prochains mois.

Diplomatie de l'alignement : la chaîne matérielle comme axe de souveraineté

Alors que l'IA modifie singulièrement les rapports de force internationaux, l'un des enjeux majeurs qui animent actuellement l'écosystème international réside dans la détermination d'une instance transnationale dédiée à la gouvernance de l'intelligence artificielle.

À cet égard, la France s'est positionnée sur la recherche et les applications techniques de l'IA, en lançant, avec le Canada, le Partenariat mondial pour l'IA (PMIA) en 2020, dont le secrétariat est assuré à Paris par l'OCDE. À l'occasion du Sommet pour l'action sur l'intelligence artificielle, organisé à Paris en février 2025, **la déclaration de Paris a promu, quant à elle, la conception de systèmes d'IA inclusifs et durables, renforçant l'approche éthique et la notion d'IA de confiance**, prenant ses distances avec l'influence long-termiste et d'altruisme effectif du Sommet de Londres de novembre 2023. L'adoption du Pacte numérique mondial par l'ONU, dans le même temps, promeut une gouvernance de l'IA plus inclusive et plus efficace.

Les groupes scientifiques indépendants créés par l'ONU et la Commission européenne, constituent une nouvelle forme de gouvernance axée sur la mise en cohérence des savoirs et la hiérarchisation des risques, avant même de réglementer.

La vision française de l'IA, centrée autour de l'éthique, de la frugalité, et de la mutualisation, participe de l'image de marque différenciée dont nous disposons dans d'autres secteurs : la haute couture plutôt que la mode éphémère, la gastronomie plutôt que la malbouffe, et l'intelligence robuste plutôt que la performance. **Face au gigantisme, cette approche mérite d'être défendue dans les différents *forums* internationaux et est adaptée aux contextes à infrastructure limitée.**

Recommandation n°39 (MT) : **Publier une doctrine nationale « Smart IA » : un usage intelligent et élégant de l'IA**

(a) Accompagner le changement de paradigme à l'échelle des utilisateurs, afin de mieux qualifier et distribuer les usages : « *la bonne IA au bon endroit* ». Cette réflexion inclut notamment l'économie de la cognition, les risques psychosociaux, la frugalité, et l'éthique ; et doit prolonger le référentiel français pour une IA frugale publié par l'Écolab et l'ADEME, l'outil ComparIA du Ministère de la Culture -déjà utilisé par d'autres pays européens-, et le projet CAUSALI-T-AI piloté par le CEA, le CNRS, et l'Inria, dans le cadre du PEPR IA ;

(b) Soutenir les projets d'IA alternatifs :

généralistes : 29 experts de 10 pays différents, sous l'égide de Yoshua Bengio, appellent à mutualiser capital, compétences, données, et infrastructures de calcul en vue de développer des modèles de frontière[174] ;

physiques : Yann Le Cun, fondateur d'AMI Labs, défend une confédération d'États souverains autour d'un stack mondial, ouverte et adaptable aux besoins uniques de chaque pays dans son projet « Tapestry », lancé en avril 2026 avec la coalition AI Alliance.

[174] « À Blueprint for Multinational Advanced AI Development », note co-rédigée par 29 experts : Adrien Abecassis, Jonathan Barry, Ima Bello, Yoshua Bengio, Antonin Bergeaud, Yann Bonnet, Philipp Hacker, Ben Harack, Sophia Hatz, Joachim Henkel, Holger H. Hoos, Kit Kitamura, Ranjit Lall, Yann Lechelle, Constance de Leusse, Charles Martinet, Nicolas Miailhe, Julia C. Morse, Maximilian Negele, Kyung Ryul Park, Miro Pluckebaum, Murielle Popa-Fabre, Benjamin Prud'homme, Yohann Ralle, Mark Robinson, Charbel-Raphael Segerie, José-Ignacio Torreblanca, Lucia Velasco, K. VijayRaghavan, novembre 2025

Le modèle actuel de compétition privée et de rivalité inter-étatique amplifie les risques, et la concurrence normative s'intensifie : le *NIST AI Risk Mangagement Framework* est déployé activement par l'USAID dans ses programmes de coopération en Afrique[175], tandis que la Chine a officialisé WAICO, une alliance de 29 pays – aucun occidental-, le 16 juillet 2026, et dont le siège sera à Shanghai.

> « *En Chine, on dit souvent qu'une seule corde ne suffit pas à faire de la musique et qu'un seul arbre ne fait pas une forêt. Le développement de l'IA ne doit pas être une performance solo d'un unique pays mais une symphonie de coopération internationale.* »
>
> Xi Jinping[176]

Si la France ne peut, à raison, agir seule, il existe un marché international pour des technologies résilientes à la dépendance américaine ou chinoise, qu'elle peut adresser.

La fenêtre pour une action coordonnée et déterminée franco-européenne est étroite : dans 2 à 5 ans, les pays auront vraisemblablement fait leurs choix d'alignement réglementaire et industriel. Les pays tiers pourraient préférer s'aligner sur les règles sino-américaines, plus simples et soutenues par des marchés plus vastes, plutôt que sur le modèle européen, perçu comme trop rigide.[177]

Sous l'impulsion de la France, l'Union Européenne a tout intérêt à porter des engagements vérifiables (« *if-then* ») communs sur les capacités de frontière, exportables *via* les enceintes existantes (réseau des *AI Safety Institutes*, OCDE, ONU), mais aussi de manière plus informelle avec les pays qui conçoivent la gouvernance de l'IA d'une manière analogue.

À titre d'exemple, l'Inde et sa vision systémique d'« *AI for public good* », ainsi que le Canada et sa stratégie « IA pour tous »[178], sont des pays avec qui la France peut porter les thématiques de l'alignement : robustesse, transparence, confiance.

Recommandation n°40 (MT) : **Lancer et prendre la tête d'une « coalition des alignés »** pour matérialiser la 3e voie des pays non-alignés avec la trajectoire sino-américaine de gouvernance de l'IA

(a) Rassembler, sur le modèle de la « Coalition des Volontaires »[179], les pays attachés à la prise en compte des valeurs démocratiques et de l'éthique dans la gouvernance de l'IA ;

[175] National Institute of Standards and Technology (NIST), AI Risk Management Framework, mise à jour de mars 2025, U.S. Department of Commerce; United States Agency for International Development (USAID), AI in Global Development Playbook, janvier 2024.

[176] Discours de Xi Jinping lors de l'ouverture de la Conférence mondiale sur l'IA à Shanghai, officialisant WAICO, le 17 juillet 2026, en présence d'Antonio Guterres, secrétaire général de l'ONU.

[177] ECI Chine

[178] ECI Canada

[179] Lancée par la France et le Royaume-Uni le 17 février 2025 pour garantir une paix durable en Ukraine après la fin des hostilités.

(b) Structurer un forum de coopération francophone pour porter des positions communes dans les enceintes de normalisation internationales (ISO, UIT) et co-construire des référentiels, volontaires et complémentaires, pour les domaines non couverts par le RIA (éducation, agriculture, services publics non critiques…), et intégrant explicitement les contraintes des contextes à infrastructure numérique limitée : sobriété énergétique, modularité, simplicité des dispositifs d'évaluation, multilinguisme. Le LNE et l'AFNOR pourraient représenter la France et exercer le premier secrétariat tournant, et associer la CIRAD, l'AFD, l'INRAE, Expertise France… ;

(c) Conclure des conventions bilatérales avec des pays partenaires prioritaires (Tunisie, Maroc), et adosser le dispositif à des bourses et chaires d'accueil dans les écoles françaises. L'objectif serait d'accompagner ces pays dans la construction de leur propre cadre réglementaire IA en s'inspirant du modèle européen, tout en respectant leurs choix souverains, et sur la base de la stratégie européenne *Global Gateway*[180] ;

Toutefois, le maintien du dialogue avec les États-Unis (qui assureront la présidence du G7 en 2027) et la Chine est inévitable, à des fins de coordination des risques émanant des modèles à la frontière (GPAI) et des modèles *open-weight* ;

(d) Organiser des séminaires conjoints et des missions d'étude sur un mode de coopération plus souple.

L'ensemble de ces recommandations supposent une action diplomatique proactive, mobilisant la Représentation permanente auprès de l'UE pour une articulation fine avec le droit de l'Union, ainsi que la constitution d'un noyau dur avec l'Allemagne, qui compte également des pépites en IA (Black Forest Labs, Aleph Alpha, Deepl). Dans la continuité du sommet de Berlin de novembre 2025, le consortium franco-allemand, sous statut d'EDIC, AI Frontier Lab est un outil juridique pertinent (*cf supra*, p.18).

La France, et l'Europe, sous l'étendard de la BITN-IA, peuvent articuler robustesse, frugalité, éthique, et diversité, et les affirmer comme troncs de cette « forêt intelligente » qui dégage une troisième voie, ni descendante, ni ascendante, mais transcendante.

[180] Lancée en 2021 sous Présidence française du Conseil de l'UE, la stratégie Global Gatway prévoit de mobiliser 300 Mds€ d'ici 2027 pour soutenir le financement d'infrastructures, notamment dans le numérique, dans les pays émergents, et prioritairement sur le continent africain.

Partie III

Dix priorités pour passer de la doctrine à l'action

Les recommandations formulées dans le présent rapport dessinent une politique d'ensemble. Elles ne peuvent être réduites à une succession de mesures indépendantes : leur efficacité dépend de leur articulation autour d'une doctrine commune, de capacités scientifiques et techniques, d'une gouvernance opérationnelle, d'une politique industrielle et d'une stratégie de résilience.

Afin d'en faciliter l'appropriation et la mise en œuvre, elles sont regroupées en dix priorités d'action.

Pour chacune de ces priorités, deux horizons sont distingués :

- **Décider maintenant :** les décisions pouvant être engagées à court terme à partir de compétences, d'institutions ou de leviers déjà disponibles ;
- **Construire dans la durée :** les transformations nécessitant une montée en capacité, des investissements pluriannuels, une évolution législative ou une action européenne ou internationale.

Les vecteurs juridiques courts (circulaire, décret, loi de finances) permettent une exécution rapide, sous réserve que la volonté politique accompagne la durée nécessaire à la transformation effective. Cette distinction ne signifie pas que les actions décidées immédiatement produiront toutes leurs conséquences à court terme. Plusieurs programmes scientifiques, industriels ou institutionnels doivent être engagés sans délai précisément parce que leur mise en œuvre nécessitera plusieurs années.

Une feuille de route cohérente et progressive

À court terme, la France peut clarifier sa doctrine, consolider l'INESIA, renforcer les autorités sectorielles, structurer les parcours d'accompagnement, utiliser la commande publique, lancer les programmes de recherche et préparer les exercices de résilience.

Dans la durée, elle doit porter au niveau européen et international les conditions d'une capacité autonome d'évaluation, d'une infrastructure de preuve, d'une filière industrielle de l'IA alignée, d'une gouvernance des modèles de frontière et d'une coopération organisée face aux risques systémiques.

Ces deux horizons sont indissociables. Les décisions immédiates donnent une réalité opérationnelle à la doctrine ; les transformations structurelles permettent de tenir dans la durée face à l'évolution rapide des capacités des systèmes d'intelligence artificielle.

Priorité n° 1

Porter une doctrine française et européenne commune de l'alignement

Objectif — Stabiliser une définition partagée de l'alignement et traduire cette doctrine en exigences proportionnées à la performance des systèmes, à la criticité de leurs usages et au caractère réversible ou irréversible de leur diffusion.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Porter une définition commune de l'alignement au niveau européen. Inscrire durablement l'alignement dans l'agenda européen et faire converger les travaux des autorités nationales autour d'une définition commune. *Référence : Recommandation n° 1*	Premier ministre et ministre chargé de l'Intelligence artificielle et du Numérique, avec le MEAE, la DGE, le CIANum et l'INESIA	Position française ; initiative auprès de la Commission et des États membres ; Conseil de l'Union européenne ; proposition de résolution européenne ; réseau des AI Safety Institutes
Anticiper la mise en conformité par les engagements volontaires. Promouvoir l'adhésion des opérateurs français à l'AI Pact et aux chartes professionnelles anticipant l'application complète du RIA. *Référence : Recommandation n° 26*	Ministre chargé de l'Intelligence artificielle et du Numérique et DGE, avec les organisations professionnelles	AI Pact ; chartes professionnelles ; mobilisation des filières ; engagements volontaires ; communication gouvernementale

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Définir des exigences proportionnées à la performance et à la criticité. Articuler l'approche par les risques prévue par le RIA avec une appréciation intrinsèque des capacités des modèles et des conséquences de leur déploiement. *Référence : Orientation structurante du rapport*	DGE et INESIA, avec les autorités sectorielles, le CIANum et l'AI Office	Doctrine nationale ; mise en œuvre du RIA ; référentiels d'évaluation ; normalisation européenne
Définir un seuil de capacité pour la mise à disposition des poids des modèles. Maintenir leur libre diffusion en deçà d'un seuil et la conditionner, au-delà, à une évaluation préalable des risques irréversibles. *Référence : Recommandation n° 21*	Ministre chargé de l'Intelligence artificielle et du Numérique et DGE, avec l'INESIA et l'AI Office	Évolution du cadre européen ; critères de capacité ; benchmarks ; travaux de la Commission européenne

Priorité n° 2

Placer l'humain, l'éthique et la littératie au cœur des usages

Objectif — Faire de la participation des utilisateurs, de la délibération éthique, de la supervision humaine et de la compréhension des systèmes des composantes constitutives de l'alignement.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Accompagner la réflexion éthique sur l'ensemble du cycle de vie des SIA. Systématiser la délibération éthique avant leur déploiement dans un service public, associer les parties prenantes et mettre en place des référents à l'éthique. *Référence : Recommandation n° 4*	Premier ministre, Ariane et DITP, avec les ministères, les collectivités, les instances de dialogue social et les représentants des usagers	Doctrine interministérielle ; instructions ; référentiels éthiques ; organisation des services ; dialogue social ; chartes d'usage ; décret modifiant les conditions d'accès aux emplois fonctionnels avec formation IA obligatoire.
Opérationnaliser les valeurs par les usages et les bonnes pratiques. Soutenir les RegTech et promouvoir des cas d'usage documentés et des obligations lisibles. *Référence : Recommandation n° 9*	DGE, avec Ariane, la DITP, les autorités sectorielles et l'écosystème des RegTech	AMI LegalTech ; guides de bonnes pratiques ; documentation des cas d'usage ; montée en puissance du programme orchestrateurs (objectif 2030 : 10 000 formés).
Renforcer la pédagogie et l'acculturation. Intégrer les enjeux de sécurité, d'éthique et de responsabilité dans les dispositifs de sensibilisation et de formation. *Référence : Recommandation n° 10*	Ministre chargé de l'Intelligence artificielle et du Numérique, DGE et DITP, avec les employeurs publics et privés	Cafés IA ; Osez l'IA ; formation professionnelle ; certifications ; modules pédagogiques
Garantir le droit effectif au recours. Maintenir un accompagnement non numérique de qualité équivalente pour tout service public dématérialisé. *Référence : Recommandation n°28*	Premier ministre, Ariane et DITP, avec France Services, les collectivités, les associations d'élus et le Défenseur des droits	Doctrine de qualité du service public ; organisation des services ; conventions avec les collectivités ; cahiers des charges

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Lutter contre l'anthropomorphisation des modèles d'IA. Publier leurs spécifications comportementales, prévoir un avertissement aux utilisateurs, encadrer l'usage du « je » et promouvoir les systèmes de recommandation optimisant le bien commun. *Référence : Recommandation n° 5*	Ministre chargé de l'Intelligence artificielle et du Numérique, avec la DGE, la DGCCRF, l'ARCOM, la CNIL et l'AI Office	Lignes directrices ; obligations de transparence ; droit de la consommation ; RIA ; codes de bonnes pratiques
Instaurer un service public de l'algorithme dans certains domaines sensibles. Évaluer l'impact des modèles au regard des objectifs de service public et privilégier, lorsque cela est nécessaire, des modèles maîtrisés et adaptés aux finalités éducatives. *Référence : Recommandation n° 17*	Premier ministre et Ariane, avec les ministères concernés, l'INESIA, le PEReN et les collectivités	Organisation du service public ; infrastructures publiques ; commande publique ; protocoles d'évaluation ; curation des données

Priorité n° 3

Faire de l'interprétabilité et de la contrôlabilité une priorité de recherche

Objectif — Rééquilibrer l'effort de recherche consacré à l'augmentation des capacités des modèles en investissant dans la compréhension, la contrôlabilité et la vérification de leur fonctionnement.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Lancer un Grand Défi européen de l'interprétabilité et de la contrôlabilité. Organiser un programme de 24 à 36 mois associant red team indépendante, équipes concurrentes, calcul dédié et pile logicielle d'audit ouverte. *Référence : Recommandation n° 3*	Premier ministre et SGPI, avec le ministère chargé de la Recherche, l'Inria, le CNRS, le CEA et l'INESIA	Programme de recherche ; appel à projets ; challenge scientifique ; financement pluriannuel ; capacités de calcul
Investir dans la recherche appliquée de la vérification formelle. Développer les méthodes formelles, la cryptographie, la sécurité matérielle et le contrôle des environnements d'entraînement. *Référence : Recommandation n° 14*	Ministère chargé de la Recherche et Inria, avec le CEA, l'ANSSI, l'INESIA et les laboratoires concernés	Programme Apollo ; programmation de la recherche ; financements nationaux et européens ; partenariats public-privé
Structurer une plateforme ouverte et multicentrique d'évaluation. Réduire les délais, organiser des défis, définir des livrables et mutualiser les actifs technologiques. *Référence : Recommandation n° 22*	Ministère chargé de la Recherche, avec le CNRS, le CEA, l'Inria et l'INESIA	Plateforme collaborative ; simplification contractuelle ; challenges ; mutualisation des brevets et actifs

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Créer une ARPA européenne dédiée à l'IA sûre par conception. Financer les projets à haut risque et fort potentiel dans les domaines des preuves formelles, de la vérification matérielle, de l'interprétabilité et des architectures contrôlables. *Référence : Recommandation n° 2*	France et Commission européenne, avec l'AI Office, l'INESIA et les organismes de recherche	Initiative européenne ; budget européen ; financements high risk / high reward ; certification ; label européen

Priorité n° 4

Construire une capacité indépendante d'évaluation et de vérification

Objectif — Donner aux autorités publiques et aux chercheurs les moyens d'accéder aux modèles, d'évaluer leurs capacités et de vérifier les garanties avancées avant et après leur déploiement.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Doter la recherche publique d'infrastructures indépendantes et sécurisées. Créer des enclaves d'évaluation confidentielles et des capacités de calcul classifiées. *Référence : Recommandation n° 13*	SGDSN et INESIA, avec l'Inria, le MINARM (AMIAD), l'ANSSI, le LNE, le PEReN et les opérateurs de calcul	Infrastructures sécurisées ; conventions d'accès ; capacités classifiées ; supercalculateur Asgard ; financement public
Financer une capacité européenne de post-entraînement. Ré-aligner les modèles à poids, rechercher les portes dérobées et les évaluer dans leur contexte de déploiement. *Référence : Recommandation n° 18*	INESIA et Inria, avec le CEA, les opérateurs de calcul et les partenaires européens	Accès contractuel aux modèles ; programmes de post-entraînement ; capacités de calcul ; partenariats de recherche

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Rendre obligatoire la validation expérimentale des modèles de frontière. Prévoir red teaming indépendant, évaluations reproductibles, seuils de capacités et signalement des incidents. *Référence : Recommandation n° 16*	AI Office et autorités nationales compétentes, avec l'INESIA et la DGCCRF	Cadre européen ; accès au marché ; protocoles standardisés ; obligations de signalement ; engagements « if-then »
Adosser un institut européen d'évaluation à l'AI Office. Le doter de calcul, de personnels spécialisés, d'un accès aux modèles et d'une capacité d'exécution autonome. *Référence : Recommandation n° 24*	Commission européenne et AI Office, en articulation avec les instituts nationaux	Initiative européenne ; budget de l'Union ; évolution du mandat de l'AI Office ; accords d'accès
Structurer l'écosystème européen de spécification-vérification. Coordonner les compétences européennes et développer des mécanismes matériels de vérification. *Référence : Recommandation n° 35*	AI Office, avec l'Inria, l'INESIA, les organismes de recherche et les industriels	Réseau européen ; normalisation ; partenariats industriels ; accords internationaux ; vérification matérielle

Priorité n° 5

Bâtir une infrastructure de preuve, de traçabilité et d'auditabilité

Objectif — Passer d'une conformité essentiellement déclarative à des garanties fondées sur des faits vérifiables, des traces opposables et l'observation du comportement réel des systèmes.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Articuler les obligations du RIA avec une exigence de vérifiabilité technique. Relier les obligations de documentation et de transparence aux artefacts techniques permettant de les vérifier. *Référence : Recommandation n° 6*	DGE, Conseil d'État, CNIL, ARCOM et DGCCRF, avec le PEReN et l'INESIA	Doctrine d'application du RIA ; habilitations juridiques ; référentiels techniques ; obligations de documentation
Intégrer l'orchestration dans les référentiels d'évaluation. Inclure les standards BPMN et DMN afin que les systèmes agentiques opèrent dans un cadre auditable. *Référence : Recommandation n° 7*	DGE et AFNOR, avec l'INESIA, le PEReN et les organismes européens de normalisation	Référentiels d'évaluation ; normalisation ISO et européenne ; cahiers des charges ; standards ouverts
Rendre observables les engagements des fournisseurs. Publier leurs spécifications comportementales et leurs politiques de sécurité ainsi que les seuils et mesures associés. *Référence : Recommandation n°5 et 32*	DGE, DGCCRF et INESIA, avec l'AI Office	Transparence ; obligations contractuelles ; RIA ; codes de bonnes pratiques ; conditions d'accès au marché

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Soutenir des standards ouverts de traçabilité et de preuve. Structurer une filière européenne de l'infrastructure probatoire et un comité de pilotage associant normalisation, recherche, industrie et régulateurs. *Référence : Recommandation n° 8*	AFNOR et DGE, avec le CEN-CENELEC, l'Inria, les autorités de régulation et les secteurs utilisateurs	Normalisation française et européenne ; standards ouverts ; comité de pilotage ; reconnaissance juridique des mécanismes de preuve

Priorité n° 6

Doter la France d'une gouvernance opérationnelle de l'alignement

Objectif — Consolider les institutions existantes, clarifier leurs responsabilités et leur donner les moyens humains, scientifiques et techniques nécessaires pour agir au rythme des capacités de l'IA.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Établir une cartographie dynamique de l'écosystème. Recenser les compétences et les acteurs publics, privés et académiques de l'alignement en France puis en Europe. *Référence : Recommandation n° 22*	CIANum, avec le CNRS, Ariane, la DGE et l'INESIA	Lettre de mission ; portail national ; recensement dynamique
Doter l'INESIA d'un budget propre et de moyens humains dédiés. Lui conférer un statut juridique, un budget pluriannuel, un secrétariat général permanent et un pôle de prospective. *Référence : Recommandation n° 23*	Premier ministre et SGDSN, avec la DGE et les ministères de tutelle	Arbitrage interministériel ; texte réglementaire ; programmation budgétaire ; recrutements hors grille
Renforcer les capacités des autorités sectorielles. Organiser l'appui du PEReN, diffuser une doctrine commune, mutualiser les modèles auditables et préparer un rescrit CNIL-INESIA. *Référence : Recommandation n° 25*	DGE, avec le PEReN, l'INESIA, la CNIL, la DGCCRF et les autorités sectorielles	Doctrine administrative ; mutualisation ; guichet commun ; rescrit ; accords de reconnaissance mutuelle
Créer des pôles d'expertise sectoriels au sein de l'INESIA. Élaborer des méthodologies, benchmarks et services de certification pour les secteurs prioritaires. *Référence : Recommandation n° 36*	INESIA et SGDSN, avec les autorités, chercheurs et industriels concernés	Organisation de l'INESIA ; programmes sectoriels ; partenariats ; certification

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Valider par la loi la désignation des autorités compétentes. Assurer la lisibilité du dispositif national et habiliter la CNIL à participer au mécanisme de rescrit. *Référence : Recommandation n°25*	Gouvernement et Parlement, avec la DGE, la DGCCRF et la CNIL	Projet de loi ; débat parlementaire ; habilitations législatives ; désignation des autorités

Priorité n° 7

Transformer la conformité en accompagnement et en avantage compétitif

Objectif — Donner aux entreprises des parcours lisibles de mise en conformité, développer l'offre d'audit et de certification et faire des exigences européennes un facteur de confiance, d'adoption et de différenciation.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Flécher un parcours d'accompagnement des entreprises. Associer évaluations, red teaming, tests sectoriels, audits indépendants, certifications et traitement des incidents. *Référence : Recommandation n° 12*	DGE et INESIA, avec Bpifrance, les autorités sectorielles, les évaluateurs et organismes de certification	Dispositifs d'accompagnement ; appels à projets ; audits accrédités ; référentiels ; certifications ; introduire une différenciation graduée des obligations selon la taille de l'entreprise lors de la révision du RIA en 2028.
Intégrer l'alignement dans les relations contractuelles. Introduire des exigences dans les CCAG et faciliter l'interopérabilité par des critères minimaux et des chartes. *Référence : Recommandation n° 19*	DAJ et DGE, avec Ariane, les acheteurs publics et les organisations professionnelles	CCAG ; clauses-types ; chartes ; protocoles d'interopérabilité ; critères extra-financiers
Rendre effectifs les bacs à sable réglementaires. Garantir leur accessibilité aux PME et ETI et organiser la coopération entre autorités. *Référence : Recommandation n° 20*	DGE, avec la CNIL, l'ARCOM, l'AMF, l'ACPR et l'INESIA	Bacs à sable du RIA ; coopération inter-autorités ; accompagnement réglementaire ; expérimentation, décret PM porté avec CNIL et INESIA
Renforcer la présence française dans la normalisation. Mobiliser les filières, accélérer la désignation d'organismes notifiés et porter un label européen. *Référence : Recommandation n° 26*	DGE et AFNOR, avec le LNE, le Cofrac et les fédérations professionnelles	Normalisation ; accréditation ; organismes notifiés ; label européen ; coordination des filières

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Structurer un écosystème assurantiel de l'alignement. Définir l'incident d'alignement, organiser sa notification et moduler les primes selon la maturité des pratiques. *Référence : Recommandation n° 11*	Ministère de l'Économie et des Finances, avec l'ACPR, l'AMF, l'INESIA et les assureurs	Définition juridique ; notification ; audits indépendants ; notation des pratiques ; mécanisme assurantiel
Adapter les certifications sectorielles aux composants apprenants. Faire évoluer les cadres applicables au nucléaire, à l'aéronautique, à l'automobile, au médical et à la sûreté industrielle. *Référence : Recommandation n° 37*	Autorités sectorielles et organismes de certification, avec l'INESIA, le LNE, le CEA et les industriels	Normes sectorielles ; certification ; recherche appliquée ; coopération européenne et internationale

Priorité n° 8

Faire de l'État et de la commande un moteur de l'IA alignée

Objectif — Utiliser les achats, les infrastructures et les usages publics pour structurer la demande, diffuser des exigences vérifiables et éviter la duplication de solutions coûteuses ou insuffisamment maîtrisées.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Mutualiser les moyens et les solutions des administrations. Créer un catalogue national de cas d'usage et de briques réutilisables et mutualiser l'accès aux modèles auditables. *Référence : Recommandation n° 15*	Ariane, avec les ministères, les autorités indépendantes, les collectivités et l'INESIA	Catalogue national ; mutualisation interministérielle ; modèles auditables ; conventions avec les collectivités
Élaborer des clauses-types d'achat de systèmes d'IA. Intégrer des critères de transparence, maîtrise des données, évaluation, portabilité, réversibilité et sobriété. *Référence : Recommandations nos 19 et 31*	Premier ministre, DAJ et Ariane, avec la DGE, la DITP, la DGFiP, l'ADEME et les acheteurs	Clauses-types ; CCAG ; cahiers des charges ; politique d'achat ; référentiels d'écoconception et de résilience
Objectiver les critères d'alignement exigibles par les acheteurs. Formaliser des notations comparables, rendre les engagements vérifiables et imposer la portabilité. *Référence : Recommandation n° 32*	DGE et DAJ, avec l'INESIA, les acheteurs et les tiers auditeurs	Notation ; audits indépendants ; clauses contractuelles ; politiques de sécurité ; critères de sélection
Faire de la sobriété numérique un critère de déploiement. Intégrer l'évaluation environnementale et publier l'empreinte carbone des systèmes publics. *Référence : Recommandation n° 38*	Premier ministre, ADEME, Ariane, DITP, DAJ et Commissariat général au développement durable	Cahiers des charges ; évaluation environnementale ; méthodologie ADEME-Ariane ; publication annuelle ; clauses-types

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Faire de la commande publique et privée un levier de structuration de la filière. Orienter la recherche partenariale, promouvoir la préférence européenne et prévoir des clauses de substitution ou de réversibilité. *Référence : Recommandation n° 31*	Premier ministre et ministère de l'Économie et des Finances, avec la DGE, le SGPI, les grands donneurs d'ordre et les institutions européennes	Commande publique et privée ; recherche partenariale ; financement d'infrastructures ; initiative européenne ; clauses de substitution

Priorité n° 9

Structurer une base industrielle et technologique de l'IA alignée

Objectif — Concentrer l'effort public et privé sur les briques rares et créatrices de valeur et maîtriser les conditions d'accès aux ressources critiques.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Publier une doctrine nationale « Smart IA ». Promouvoir « la bonne IA au bon endroit » en intégrant frugalité, éthique, risques psychosociaux et économie de la cognition. *Référence : Recommandation n° 39*	Premier ministre et ministre chargé de l'Intelligence artificielle et du Numérique, avec Ariane, la DITP, l'ADEME et le CIANum, MEAE, Ministère de la Culture	Doctrine nationale ; guides d'usage ; communication gouvernementale ; politiques de formation et d'achat ; publication du référentiel français volontaire « IA frugale ».
Créer une offre européenne intégrée sous une bannière commune. Associer capacités de calcul, cloud, énergie, modèles et briques logicielles et faciliter les consortiums. *Référence : Recommandation n° 33*	DGE et MEAE, avec les industriels, les ambassadeurs et les conseillers du commerce extérieur	Consortiums ; offres groupées ; diplomatie économique ; missions des ambassades ; coopération industrielle européenne

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Consolider une BITN-IA. Agréger les moyens publics, industriels, scientifiques et financiers et mobiliser les financements nécessaires au passage à l'échelle. *Référence : Recommandation n°30*	Premier ministre et ministère de l'Économie et des Finances, avec la DGE, le SGPI, la CDC, Bpifrance, les investisseurs et les industriels	Fonds d'infrastructure ; préachats ; avances remboursables ; France 2030 ; initiative Tibi ; FEI ; projet de loi de finances : crédit d'impôt recherche reventilé avec mécanisme correcteur, abondement pour la mise en conformité (Bpifrance, France 2030)
Pondérer l'accès aux ressources et aux capacités de calcul. Prévoir un droit de regard, réserver une part des capacités aux acteurs européens, conditionner l'accès à certaines données et soumettre les modèles les plus avancés à une autorisation de mise sur le marché. *Référence : Recommandation n° 34*	Premier ministre et ministère de l'Économie et des Finances, avec les ministères chargés de l'Énergie, de l'Industrie et du Numérique, l'INESIA et les collectivités	Allocation du foncier et de l'énergie ; conditions de financement ; droits d'accès ; conventions ; réglementation nationale et européenne ; autorisation de mise sur le marché ; convention GENCI-Etat pour l'accès de Jean Zay à des startups (portée avec MESRI et DGE)

Priorité n° 10

Préparer la résilience et porter une diplomatie de l'alignement

Objectif — Anticiper la défaillance des garanties, préparer la continuité des fonctions essentielles, organiser la reprise de maîtrise et porter une coopération internationale adaptée.

Décider maintenant

Action principale	Pilotage proposé	Leviers
Spécialiser l'INESIA sur les risques systémiques. Orienter l'effort vers l'IA défensive et publier un tableau de bord de la dépendance cognitive et technologique. *Référence : Recommandation n° 27*	Premier ministre et SGDSN, avec l'INESIA, l'ANSSI et les administrations concernées	Feuille de route de l'INESIA ; financement de l'IA défensive ; tableau de bord public ; cartographie des dépendances
Institutionnaliser des exercices de perte de maîtrise. Tester l'exfiltration de poids, les comportements agentiques hors périmètre et la compromission d'infrastructures critiques. *Référence : Recommandation n° 28*	Premier ministre et SGDSN, avec l'INESIA, l'ANSSI, les OIV, les autorités sectorielles et les partenaires européens	Exercices de crise ; scénarios types ; base d'incidents ; retours d'expérience ; révision de la doctrine
Garantir des capacités de fonctionnement sans IA. Maintenir des procédures permettant aux OIV d'assurer leurs missions essentielles en cas d'interruption. *Référence : Recommandation n°28*	SGDSN et ministères sectoriels, avec les OIV et les autorités de régulation	Plans de continuité ; obligations sectorielles ; exercices ; procédures de mode dégradé ; contrôle des opérateurs
Constituer une contre-expertise indépendante sur les risques NRBC-e et les usages duaux. Associer autorités de sûreté et experts scientifiques compétents. *Référence : Recommandation n° 36*	SGDSN et INESIA, avec l'ASNR, les autorités sectorielles et les organismes de recherche	Expertise scientifique ; protocoles d'évaluation ; coopération avec les autorités de sûreté ; accès sécurisé aux modèles

Construire dans la durée

Action structurante	Pilotage proposé	Leviers
Porter une instance internationale de coordination de la résilience. Organiser la coordination technique, la notification urgente des incidents et la supervision des normes. *Référence : Recommandation n° 29*	Président de la République, Premier ministre et MEAE, avec le SGDSN et l'INESIA	Initiative diplomatique ; G7 ; OCDE ; ONU ; réseau des AI Safety Institutes ; coopération internationale
Lancer une « coalition des alignés ». Rassembler les pays attachés aux valeurs démocratiques, à l'éthique, à la robustesse et à la transparence et soutenir des projets mutualisés. *Référence : Recommandation n° 40*	Président de la République et MEAE, avec le ministre chargé de l'IA, l'INESIA et les partenaires internationaux	Coalition interétatique ; partenariats de recherche ; diplomatie scientifique ; OCDE ; ONU ; G7 ; réseau des AI Safety Institutes

Conclusion générale

Aligner, c'est s'adapter en continu

Après nous être demandé s'il y avait un pilote dans l'avion, « y a-t-il un harnais dans le système d'IA ? » pourrait être notre prochain questionnement.

L'alignement peut, en effet, être pour l'Europe ce que la sûreté a été pour son industrie aéronautique : une exigence devenue avantage compétitif mondial, que nul ne remet plus en cause aujourd'hui.

Les développements attendus dans le champ de l'IA n'ont pas toujours été à la hauteur des attentes, tant scientifiques qu'économiques, engendrant deux périodes dites des « hivers de l'IA »[181], vectrices d'une perte d'engouement et donc de financements sur le sujet, tant pour la recherche scientifique, que pour les entreprises innovantes.

Pour que l'alignement devienne le « nouveau printemps de l'IA », la fenêtre d'opportunité est étroite, à mesure que le marché de la confiance marginalise l'Europe, que les États-Unis et la Chine carburent à la performance, et que les fournisseurs de modèles relèguent l'Europe, à tort – parce que cela les arrange bien –, dans une surrégulation anachronique.

Rendre l'alignement opérationnel, c'est d'abord redonner à la délibération publique la place qui lui revient dans des décisions qui, sous couvert de technique ou de performance, sont toujours, aussi, des choix de société.

L'alignement, comme le décrit ce rapport, est une propriété produite par une succession d'opérations, conditionnant la fiabilité du SIA à la maitrise de bout en bout des systèmes, des usages, des dépendances, des chocs. La régulation adaptative, que j'appelle de mes vœux, ne signifie pas l'absence de règles, mais la capacité de l'écosystème à corriger ses déséquilibres grâce à des boucles de rétroaction distribuées. Seule cette dynamique, étoffée par des formes de coopération renouvelées, nous permettra de nous adapter aux continuels changements, sans perdre de notre cohérence.

Parce que l'alignement intégral ne peut être assuré, **notre impératif existentiel est de construire les conditions d'un alignement effectif, dynamique et vérifiable. C'est à cette condition que l'IA pourra devenir un projet collectif, industriel, démocratique et souverain pour le bien commun.**

Il me semble, à titre personnel, que nous devons le faire au service d'un intérêt général supérieur, pour une Nation et un État « augmentés » et pour que le progrès reste dans l'homme, plutôt que dans la technique[182].

---

[181] OPECST, *Les nouveaux développements de l'intelligence artificielle*, Ass. Nat., n° 642, 29 novembre 2024 : Un premier hiver a lieu au début des années 1970, puis un second dans les années 1990. Le rapport décrit une certaine cyclicité des investissements en intelligence artificielle selon une boucle « espoirs-déceptions », p.51.

[182] « *Le progrès n'est plus dans l'homme, il est dans la technique, dans le perfectionnement des méthodes capables de permettre chaque jour une utilisation plus efficace du matériel humain* », Georges Bernanos, La France et les Robots, Éditions de la France libre, 1946, p.11

« *Et la seule personne à pouvoir m'accorder une permission absolue, c'est moi-même. Une permission consciente et délibérée de voir le miracle se produire. Alors que la civilisation avance à grands pas vers la destruction absolue de la Terre, vers le suicide collectif, aveugle, de tous les êtres vivants, le dernier espoir réside dans un acte de volonté : sortir de ce piège mécanique et dire non*. »

Leonora Carrington

Annexes

Lettre de mission

Le Premier Ministre

0273/26 SG

Paris, le 27 FEV. 2026

Madame le Sénateur,

L'alignement des systèmes d'intelligence artificielle (IA) avec nos valeurs est un enjeu crucial, qui fait appel à notre capacité à mobiliser efficacement nos ressources et à embarquer nos concitoyens face à la rupture technologique incontournable que représente l'IA. Il s'agit de garantir que les systèmes d'IA puissants et autonomes restent compréhensibles, sûrs et conformes à nos valeurs.

Notre dépendance structurelle à des solutions technologiques fournies par des acteurs étrangers pourrait compromettre la confiance de nos concitoyens envers nos outils numériques, ainsi que la préservation de nos valeurs et de notre avenir collectif. L'adoption et la diffusion de l'intelligence artificielle doit être un projet collectif, industriel et sociétal, axé sur les usages et le sens de la technologie.

La France et l'Europe disposent d'un capital scientifique et humain exceptionnel, qu'il convient de valoriser pour consolider un écosystème divers et riche déjà existant et répondre à une demande nouvelle pour des systèmes d'IA « alignés » avec les notions de confiance, sécurité, durabilité et explicabilité. Il s'agit aussi de créer les conditions pour que le caractère vertueux de ces IA constitue un véritable différenciant, y compris au plan économique.

Tel était l'objet du colloque que vous avez organisé au Palais du Luxembourg, le 27 octobre 2025, qui a réuni chercheurs, scientifiques et philosophes, décideurs publics, industriels et financiers pour dessiner la stratégie qui pourrait placer la France et l'Europe à la pointe de l'alignement des IA.

Dans la lignée de la proposition de résolution européenne (PPRE), que vous avez déposée au Sénat le 27 octobre dernier pour inscrire cette thématique dans le débat européen, la structuration d'un débat sur cette question est cruciale.

.../...

Madame Vanina PAOLI-GAGIN
Sénateur
Palais du Luxembourg
15, rue de Vaugirard
75006 PARIS

Au regard de votre engagement sur ce sujet, j'ai l'honneur de vous confier une mission, qui aura vocation à :

- recenser les initiatives françaises et européennes mises en œuvre pour assurer que les systèmes d'IA agissent conformément aux intentions humaines et à nos valeurs, sur les plans de la recherche comme de la réglementation ;

- réaliser une revue de l'état de l'art en la matière : mécanismes existants pour favoriser l'alignement de l'IA, bonne pratique des développeurs, travaux internationaux de normalisation en cours ou futurs, etc. ;

- cartographier les acteurs publics, privés et académiques français et européens impliqués dans la recherche et le déploiement de solutions d'alignement de l'IA, afin de déterminer comment leur mobilisation permettrait de promouvoir, à l'international, notre expertise, nos standards et nos valeurs en matière d'alignement ;

- évaluer, sur la base de quelques cas d'usages, les conditions pour que l'IA soit pleinement alignée avec les valeurs européennes, notamment par exemple : absence de biais liés au genre, à l'origine ethnique, à la religion, etc. ;

- identifier les avantages comparatifs des solutions bien alignées avec nos valeurs pour les entreprises les développant et pour leurs utilisateurs (pertinence, rapidité et facilité d'adoption, dialogue social, etc.) ;

- sur la base de vos conclusions, formuler des recommandations à destination des développeurs de modèle ou de systèmes d'IA et des utilisateurs de ces systèmes.

En application de l'article L.O.144 du code électoral, un décret vous nommera parlementaire en mission auprès de Mme Anne LE HÉNANFF, ministre déléguée auprès du ministre de l'économie, des finances et de la souveraineté industrielle, énergétique et numérique, chargée de l'intelligence artificielle et du numérique.

Vous veillerez à élaborer vos propositions dans le respect des règles d'indépendance, d'impartialité et d'objectivité qui s'imposent au titre de la loi n° 2013-907 du 11 octobre 2013 relative à la transparence de la vie publique, et à m'informer des éventuelles mesures prises à cet effet.

Pour mener à bien cette mission, vous vous appuierez sur les services des administrations compétentes : direction générale des entreprises et direction générale du trésor. La synchronisation de ces travaux avec ceux du Conseil de l'IA et du numérique et ceux du Comité consultatif national d'éthique du numérique sera également un facteur clé de succès de votre mission.

Je souhaite pouvoir disposer de votre rapport au plus tard six mois après le début de cette mission.

Je vous prie d'agréer, Madame le Sénateur, l'expression de mes respectueux hommages.

Sébastien LECORNU

Liste des personnes entendues

I- Institutions

Secrétariat général de la Défense et de la Sécurité nationale (SGDSN)

- Nicolas Roche, secrétaire général de la Défense et de la Sécurité nationale
- Jonathan Collas, conseiller industrie et numérique
- Gwénaël Jézéquel, conseiller relations institutionnelles et communication
- Justine Boquet, chargée de mission

Service de vigilance et de protection contre les ingérences numériques étrangères (Viginum)

- Anne-Sophie Dhiver, cheffe de service adjointe

Agence nationale de la sécurité des systèmes d'information (ANSSI)

- Vincent Strubel, directeur général

Pôle d'expertise de la régulation numérique (PEReN)

- Renaud Labelle, directeur
- Victor Amblard, lead data scientist

Laboratoire national de métrologie et d'essais (LNE)

- Thomas Grenon, directeur général
- Agnès Delaborde, responsable du département évaluation de l'IA

Direction générale de la concurrence, de la consommation, et de la répression des fraudes (DGCCRF)

- Sarah Lacoche, directrice générale
- Alice Vilcot-Dutarte, directrice de cabinet
- Marion Loustric, cheffe de la cellule « Mission Data IA »

Direction générale des entreprises (DGE)

- Nina Landes, coordinatrice nationale pour l'IA

Direction générale des finances publiques (DGFiP)

- Jean-Michel Mota, directeur des projets numériques
- Khaled Bouzid, directeur du pôle données et IA

Direction interministérielle du numérique (Dinum)

- Ishan Bhojwani, responsable du département IA dans l'Etat (IAE)

Ministère des affaires étrangères

- Clara Chappaz, ambassadrice pour le numérique et l'IA, ancienne ministre déléguée chargée de l'IA et du numérique
- Lisette Rivière, chargée de mission auprès de l'ambassadrice pour le numérique et l'IA

Ministère de l'Enseignement Supérieur, de la Recherche et de l'Espace

- Jean-Luc Moullet, directeur général de la recherche et de l'innovation (DGRI)

Haut-commissariat au Plan et à la Stratégie

- Clément Beaune, haut-commissaire, ancien ministre délégué chargé de l'Europe
- Emma Coroler, cheffe de projet

Caisse des dépôts

- Patrick Laurens-Frings, directeur de la transformation opérationnelle, digitale et des SI
- Ted Marx, directeur des affaires numériques
- Nasrine Maameri, conseillère en relations institutionnelles

Agence ministérielle pour l'IA de défense (AMIAD)

- Bertrand Rondepierre, directeur
- Alexandra Le Bouhellec, responsable communication et cheffe de la cellule rayonnement
- Bénédicte Deltel, responsable de la cellule SANGRIA

Direction générale de l'armement (DGA)

- Julien Martinez-Corral, ingénieur en chef de l'armement, architecte du système de défense « numérique, cyber, quantique »
- Juliette Dubreuil, conseillère communication, relations élus, et plume auprès du DGA

Commission nationale de l'informatique et des libertés (CNIL)

- Rémi Stefanini, directeur des technologies, de l'innovation, et de l'IA
- Aurore Liboutet, chargée d'études

Association française de normalisation (AFNOR)
Branche normalisation

- Anna Médan, responsable IA
- Eric Laurencon, responsable du département électrotechnologies
- Thierry Geoffroy, responsable des affaires publiques

Branche certification

- Thomas Lommatzsch, responsable de l'unité matériel médical
- Violaine Trajan, responsable des relations institutionnelles
- Christophe Rondel, conseil

Autorité de régulation de la communication audiovisuelle et numérique (ARCOM)

- Bruno Schmutz, directeur des études et de la prospective
- Noé Beserman, chef du département données et technologies

Conseil de l'intelligence artificielle et du numérique (CIANum)

- Anne Bouverot, co-présidente, présidente de l'Ecole normale supérieure (ENS), envoyée spéciale du Président de la République pour l'IA
- Guillaume Poupard, co-président, chief trust officer chez Orange, ancien directeur général de l'ANSSI
- Joséphine Corcoral, directrice adjointe

Arno Amabile, président de l'observatoire du ministère de la Justice pour l'IA, ex-conseiller spécial de l'Envoyée spéciale pour l'IA

Comité consultatif national d'éthique du numérique (CCNEN)

- Claude Kirchner, président, directeur de recherche émérite d'Inria
- Raja Chatila, membre, professeur émérite à l'université Paris Pierre-et-Marie-Curie (UPMC), directeur de recherche au Centre national de la recherche scientifique (CNRS), président de l'Initiative mondiale IEEE sur l'éthique des systèmes autonomes et intelligents, co-président du groupe « IA responsable » du Global Partnership on Artificial Intelligence (GPAI), membre du groupe d'experts européen sur l'intelligence artificielle, ancien directeur de l'Institut des systèmes intelligents et de robotique (ISIR)
- Catherine Tessier, membre, directrice de recherche et référente intégrité scientifique et éthique à l'ONERA, membre du Comité d'éthique de la défense et du COERLE d'Inria
- Sylvie Joseph, membre, administratrice du groupe La Poste
- Mireille Régnier, membre, directrice de recherche d'Inria, présidente du Comité opérationnel d'évaluation des risques légaux et éthiques (COERLE)
- Fabien Tarissan, membre, directeur de recherche en informatique au CNRS, professeur attaché à l'ENS Paris-Saclay, membre de la CNIL

Bureau européen de l'IA (AI Office)

- Matthieu Delescluse, directeur de l'unité « sécurité de l'IA »
- Thierry Boulangé, directeur de l'unité « gouvernance et régulation »

Région Grand Est

- Raphaël Régnier, directeur du numérique
- Marc Platini, responsable des projets IA de l'agence Grand E-Nov+ (Grand Est Développement)
- Benjamin Legrand, chef de projet stratégique IA

II- Experts et scientifiques

Institut national de recherche en sciences et technologies du numérique (Inria)

- Bruno Sportisse, président-directeur général
- Jean-Michel Loubes, directeur de recherche, responsable de l'équipe « Régulation de l'IA » (REGALIA), chercheur principal de la chaire TRIAL (*Trust in Artificial Intelligence*) à l'ANITI Toulouse
- Carina Prunkl, chercheuse au sein du projet-pilote REGALIA, co-rédactrice du rapport international 2026 sur la sécurité de l'IA
- Frédérique Vidal, chargé de mission relations institutionnelles et appui aux territoires

Commissariat à l'énergie atomique et aux énergies alternatives (CEA)

- François Terrier, directeur des programmes au CEA-List, directeur du PEPR IA
- Cédric Gouy-Pailler, directeur de recherche au CEA, responsable du programme IA au CEA-LIST, référent IA pour l'agence de programme ASIC
- Alexei Grinbaum, directeur de recherche au CEA (direction de la recherche fondamentale), président du comité opérationnel pilote d'éthique du numérique du CEA, expert éthique pour la Commission européenne, membre du CCNEN
- Jérôme Bobin, directeur du département d'électronique et informatique pour la physique (Dedip) et du programme NumPEx (logiciels pour le calcul haute performance et l'IA)
- Louise Vaillant-Thomas, conseillère de l'administratrice générale du CEA

Centre national de la recherche scientifique (CNRS)

- Antoine Petit, ancien directeur général
- François Yvon, directeur de recherche au sein de l'équipe MLIA (*Machine Learning and Deep Learning for Information Access*) de l'ISIR (Institut des Systèmes Intelligents et de Robotique - CNRS/Sorbonne Université)
- Serena Villata, directrice de recherche au sein du laboratoire I3S, responsable de recherche MARIANNE au centre Inria de l'université Côte d'Azur, membre du CIANum
- Thomas Borel, responsable des affaires publiques et des relations parlementaires

Laboratoire interdisciplinaire des sciences du numérique (LISN)

- Sophie Rosset, directrice, directrice de recherche CNRS

Yoshua Bengio, professeur titulaire en informatique à l'Université de Montréal, coprésident et directeur scientifique de LoiZéro, fondateur et conseiller scientifique de Mila et titulaire d'une chaire en IA Canada-CIFAR, colauréat du Prix À.M. Turing 2018, membre du conseil scientifique consultatif de l'ONU pour les percées de la science et de la technologie, président du rapport international 2026 sur la sécurité de l'IA

David Dalrymple, ex-directeur du programme "*Safeguarded AI*" de l'Advanced Research and Innovation Agency (ARIA)

Hichem Snoussi, professeur à l'Université de Technologie de Troyes (UTT), directeur adjoint de l'unité de recherche LIST3N (Laboratoire Informatique et Société Numérique)

Alexandre Pitti, professeur à l'Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA)

Anne Alombert, philosophe, enseignante-chercheuse à l'université Paris 8 Vincennes à Saint-Denis

Jean-Gabriel Ganascia, professeur émérite à l'université Paris Pierre-et-Marie-Curie (UPMC), membre de l'Institut universitaire de France, président de l'association française pour l'avancement des sciences (Afas) et des comités d'éthique de Pôle emploi et de Docaposte

Serge Tisseron, psychiatre et docteur en psychologie, directeur de recherches à l'université Paris X, membre de l'Académie des Technologies et membre du conseil scientifique du CRPMS (Université de Paris)

Fatiha Tali Otmani, maîtresse de conférences HDR en sciences de l'éducation et de la formation à l'université Toulouse Jean Jaurès, responsable scientifique du projet EVARIATION pour la Chaire UNESCO de l'Institut national supérieur de formation et de recherche pour l'éducation inclusive (INSEI)

Mark Hunyadi, professeur de philosophie morale et politique à l'Université catholique de Louvain, membre du comité « éthique en commun » (INRAE-Cirad-Ifremer-IRD)

Nadia Guerouaou, docteure en neurosciences cognitives, chercheuse post-doctorale au Centre Internet et Société (CIS) du CNRS

Michel Sejean, professeur de droit à l'Université Sorbonne Paris Nord, co-directeur du code de la cybersécurité aux éditions Dalloz, chercheur à l'Institut de recherche pour un droit attractif (IRDA)

Brunessen Bertrand, professeure à l'université de Rennes, chaire Jean Monnet sur la gouvernance des données, membre de l'Institut universitaire de France

Marion Ho Dac, professeur de droit privé à l'université d'Artois, directrice du Centre Droit Éthique et Procédures EA 2471

Arnaud Latil, maitre de conférences HDR, vice-président IA et prospective à Sorbonne Université, membre du Sorbonne Cluster for AI, expert détaché à l'AI Office

Arnaud Touati, avocat aux barreaux de Paris et de Luxembourg en droit des nouvelles technologies, avocat associé fondateur d'Hashtag Avocats

III- Entreprises du numérique et de l'IA

Mistral AI

- Pierre Stock, vice-président des opérations scientifiques
- Cyriaque Dubois, chargé des affaires publiques
- William Ouensanga, chargé des affaires publiques

Google

- Jessica Hoffman, chercheuse chez Google DeepMind
- Sarah Boiteux, responsable des affaires gouvernementales et politiques publiques chez Google France

OpenAI

- Martin Signoux, responsable des affaires publiques

Arlequin AI

- Antoine Jardin, co-fondateur et directeur de la technologie

Probabl.ai

- Yann Lechelle, président exécutif et fondateur, membre du collectif Eurostack

Mozilla France

- Hermina Condei, directrice générale

Prophesee

- Jean Ferré, président-directeur général
- Lorraine Le Louarn, vice-présidente développement commercial
- Gautier Lagarde, responsable des relations institutionnelles

IBM

- Vincent Perrin, directeur technique écosystème, IA et informatique quantique
- Diane Dufoix-Garnier, directrice affaires publiques

Nvidia

- Ava Zekri, responsable des relations gouvernementales, France
- Lise Mustière, responsable du secteur public et défense, France

Bull

- Eric Eppe, vice-président et CTO
- Gauvain Girault

SiPearl

- Philippe Notton, président-directeur général et fondateur
- Eleanor Lasou, conseil

Vsora

- Alyssia Dauchez, cheffe de cabinet

OVH

- Blandine Eggrickx, responsable des affaires publiques
- Olivier Debeugny, fondateur et président-directeur général d'OVHai LLM (ex-Dragon LLM)

Palo Alto Networks

- Raphael Marichez, directeur sécurité Europe du Sud

Lefebvre Sarrut

- Candice Tran Dai, directrice cybersécurité

Camunda

- Stéphane Faivre-Duboz, vice-président en charge des ventes EMEA

Mirakl

- Estelle Bourgeois, responsable affaires publiques et communication

IV- Grands groupes industriels

Thales

- David Sadek, vice-président Recherche, Technologie & Innovation

Veolia

- Fouad Maach, directeur de l'Architecture, de l'Ingénierie et de l'Expertise IS&T

Framatome

- Julien Dreano, responsable de la sécurité des systèmes d'information

Airbus

- Fabrice Valentin, vice-président et directeur de l'Intelligence Artificielle, responsable de la ligne de produits et services IA de bout en bout (IA générative, Cloud).
- Adrien Ricci, directeur des affaires publiques

Alstom

- Henri Badaro, vice-président gouvernance de la data et de l'IA

Air Liquide

- Fabien Mangeant, chief data & AI officer, ex-président du comité de direction du programme confiance.ai

EDF

- Franck Michon, directeur du Centre d'Excellence Data, IA et Développement
- Martine Gouriet, directrice transformation numérique, usages et innovation
- Florent Jourde, directeur adjoint des affaires publiques

V- Secteur bancaire et financier

Autorité des marchés financiers (AMF)

- Sébastien Raspiller, secrétaire général
- Christophe Bonnet, directeur des données et de la surveillance

BNP Paribas

- Thierry Markwitz, directeur de programme DORA

Crédit Agricole

- Aldrick Zappellini, chief data & AI officer

Algoan

- Charles Ozanne, directeur des opérations

Mirova

- Guillaume Abel, directeur général délégué

Jolt Capital

- Philippe Laval, CTO et associé

Ardian

- Gonzague Boutry, responsable des infrastructures numériques Europe
- Pauline Thomson, responsable Data Science
- Thierry Caboche, responsable des affaires publiques internationales

VI- Organisations professionnelles

Impact AI

- Roxana Rugina, directrice exécutive
- Etienne Grass, membre, global chief AI officer chez Capgemini Invent, membre du CIANum
- Nathalie Beslay, membre, présidente-directrice générale de Naaia
- Guy Mamou-Mani, co-fondateur d'Open

,

Campus Cyber

- Joffrey Célestin-Urbain, président
- Jean-Noël de Galzain, président d'Hexatrust, président-directeur général de Wallix

France Digitale
- Valérian Giesz, co-fondateur de Quandela
- Nassim Ameli-Jouffroy, directrice juridique de PhotoRoom
- Jérôme Valat, président-directeur général de Cleyrop

Systematic Paris-Région
- Kamel Merzouk, coordinateur du comité de pilotage Hub Data Science & AI
- Juliette Mattioli, vice-présidente Intelligence Artificielle chez Thales, co-présidente du comité de pilotage Hub Data Science & AI
- Abdelhamid Mellouk, directeur de recherche à l'université Paris Est Créteil, co-Président du comité de pilotage Data Science & AI
- Sana Tmar, responsable de la stratégie IA chez IRT SystemX
- Sam Hee Lazare, directrice générale d'AUMOVIO
- Olivier Beaurepaire, chief Data & AI Officer, OB Conseil
- Didier Moret, président de TECHVENTURES
- Jean-Patrick Mascomere, responsable du calcul scientifique à la R&D de Total Energies
- Philippe Bereski, ex GAIA-X Projet Manager
- Matthieu Boussard, VP AI chez Craft AI

Hub France IA

- Jordan Fleurier, chargé de mission affaires réglementaires et européennes, Hub France IA
- Bertrand Cassar, vice-président du comité « IA de confiance », directeur de la valorisation stratégique des données du groupe La Poste
- Francesca Martini, co-présidente du groupe de travail « Veille », cheffe de projet « IA de confiance » au sein du groupe La Poste
- Emilie Sirvent-Hien, responsable du programme « IA responsable » chez Orange
- Arnaud Vilain, directeur enjeux d'innovation d'Orange
- Jonathan Foureur, co-fondateur de Just AI
- Marc Decombas, co-fondateur de Just AI

VII- Fournisseurs de solutions liées à l'alignement de l'IA

Giskard

- Jean-Marie John Matthews, co-fondateur et co-CEO

Numalis

- Arnault Ioualalen, président-directeur général et directeur de la recherche

KeeeX

- Laurent Henocque, fondateur et président-directeur général

Amodo Design

- Thomas Milton, co-fondateur et directeur

Quantum of Trust / Energer

- Guillaume Loizeaud, co-fondateur
- Lucie Normand, conseil

Eridanis

- Alexis Semmama, directeur général
- Pauline-Emilie Martin, responsable des relations publiques et institutionnelles

VIII- Organisations à but non lucratif

SaferAI

- Henry Papadatos, directeur général

Centre pour la Sécurité de l'IA (CeSIA)

- Charbel-Raphaël Segerie, directeur exécutif
- Pauline Charazac, responsable des relations institutionnelles

EuroStack

- Cristina Caffarra, économiste, co-fondatrice du collectif Eurostack

Joint European Disruptive Initiative (JEDI)

- André Loesekrug-Pietri, président
- Romain Forestier, directeur de recherche et chargé de projets

IX- Déplacement à Bruxelles

Lundi 23 mars 2026 - Parlement Européen

Session de travail organisée par SaferAI, « Moonshots in Reliable, Safe, and Secure AI : a path to European leverage and strategic AI adoption »

Contributions écrites

Akhrouf, Ambre et Sawadogo, Dr. Alfred, Note de position « La France et l'IA : pour un leadership nourri par la souveraineté technologique française », mai 2026, 23 pages

Benabdelkader, Romain, chercheur indépendant, Contribution écrite, avril 2026, 3 pages

CEA, Note de position, juin 2026, 7 pages

Cirad, Note « Etat des lieux et orientations stratégiques en matière d'IA », juin 2026, 3 pages

Comité consultatif national d'éthique du numérique (CCNEN), Avis n°11, « Alignement des systèmes d'IA avec les valeurs des services publics : enjeux d'éthique », juillet 2026, 62 pages

Compagnie nationale des commissaires aux comptes (CNCC), Note de position, 2 juillet 2026, 8 pages

DGCCRF, Note relative au rôle de la DGCCRF sur les questions d'alignement des SIA, juin 2026, 5 pages

Duport, Clément, président et cofondateur d'Anima Néo et PolysemIQ, Rapport de contribution à la mission sur l'alignement des systèmes d'IA, juin 2026, 20 pages

Frénod, Emmanuel, mathématicien et professeur des universités à l'Université Bretagne Sud (UBS), Note de position, mai 2026, 5 pages

Grass, Etienne, Contribution écrite pour la mission, « Où en est le problème de l'alignement des modèles d'IA », 19 mai 2026

Ho-Dac, Marion, Contribution écrite pour la mission, 12 mai 2026, 9 pages

Labelle, Renaud, PEReN, Note bibliographique, mai 2026, 10 pages

Le Cun, Yann, AMI Labs, Note de position, juillet 2026, 2 pages

Picou, Christophe, VEIA.AI, « Aligner les SIA dans les organisations : conditions opérationnelles de la souveraineté décisionnelle », mai 2026, 8 pages

Semmama, Alexis, Eridanis, Note de synthèse « L'innovation territoriale avec le projet RECITAL et la commune de Noisy-le-Grand », 1er juillet 2026, 3 pages

Touati, Arnaud, Hashtags Avocats, Contribution écrite, 10 mai 2026, 46 pages

Bibliographie

<u>*Parmi mes livres de chevet :*</u>

Adelaide, Jean-Michel, *Repenser l'intention stratégique dans un monde fluctuant, un impératif à l'ère de l'IA*, Toryo, 2026

Chervet, Flavien, *Hyperarme : Anatomie d'une intranquilité planétaire*, Nullius in Verba, 2025

Chillet, Guillaume, *Petit traité de souveraineté cognitive*, 2026

Devillers, Laurence, *Savoir vivre avec l'IA*, Denoel, 2026

Ganascia, Jean-Noel, *L'IA expliquée aux humains*, Editions du Seuil, 2024

Grallet, Guillaume, *Pionniers : voyage aux frontières de l'intelligence artificielle*, Grasset, 2025

Guerouaou, Nadia, *Notre cerveau sous influence*, Eyrolles, 2026

Hunyadi, Mark, *Déclaration universelle des droits de l'esprit humain*, PUF, 2024

Labatut, Benjamin, *Maniac*, Grasset, 2023

Matuchet, Pierre, *Révolution IA : préparer votre organisation à l'intelligence artificielle et ses agents*, 2026

Patino, Bruno, *Le temps de l'obsolescence humaine*, Grasset, 2026

Sadin, Eric, *Le désert de nous-mêmes*, L'échappée, 2025

Schaeffer, Jean-Marie, *Mythologies Web : moteurs de recherche, réseaux sociaux et intelligence artificielle*, Tracts Gallimard, n°72, 2025

Et aussi (dans l'ordre de leur apparition dans le rapport):

de la Boétie, Etienne, *Discours de la servitude volontaire*, 1576

Halpern, Gabrielle, *Tous centaures ! Éloge de l'hybridation*, PUF, 2020

Ellul, Jacques, *La technique ou l'enjeu du siècle*, Essai, 1954.

Draghi, Mario, *The Draghi Report: a competitiveness strategy for Europe*, 9 septembre 2024

SaferAI : Touzet, Chloé, Stelling, Lily, Galizzi, Bruno, « The Case for European Investment in High-Risk, High-Reward AI Reliability Research », 24 mars 2026

Bengio, Yoshua et al., *International AI Safety Report 2026*. Yoshua Bengio et al., *International AI Safety Report 2026*, publié le 3 février 2026, soutenu par plus de 30 pays, l'UE, l'OCDE et l'ONU.

Russell, Stuart, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, 2019.

Wiener, Norbert, « Some Moral and Technical Consequences of Automation », *Science* 131, 1355–1358, 6 mai 1960.

Hagendorff, Thilo, « Deception abilities emerged in large language models », *Proceedings of the National Academy of Sciences*, 121(24), juin 2024.

Lian, Jiawei, « Revealing the intrinsic ethical vulnerability of aligned large language models », *Nature Communications*, 2026.

Linder, David, et al., « Practical challenges of control monitoring in frontier AI deployments », 2025.

Kral, Simon, SaferAI, « How can Europe make AI safe and reliable? », 9 juin 2026.

SaferAI, « Unlocking European AI competitiveness and sovereignty: proposal for a European ARPA for AI RSS moonshots », 2025.

General Purpose AI Policy Lab, « Anticiper l'évolution des capacités critiques de l'IA : de la saturation des benchmarks à l'AGI », Note n° 4, janvier 2026.

CeSIA, « Vers une expertise française en évaluation de l'IA : le rôle stratégique de l'INESIA », 2025, p. 8. Voir aussi Turtayev, Rustem, et al., « Hacking CTFs with Plain Agents », 2024.

Maier, Antoine, Maier, Aude, David, Tom, « Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization », École polytechnique fédérale de Lausanne, 9 février 2026.

Pape Léon XIV, *Lettre encyclique Magnifica Humanitas*, 25 mai 2026.

Conseil d'État, étude à la demande du Premier ministre. « Intelligence artificielle et action publique : construire la confiance, servir la performance », 31 août 2022.

Shen, Hua, et al., « Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework and Future Directions », 2024.

Medrouk, Lisa, « Le premier standard qui rend le red teaming IA vérifiable est coréen », *Journal du Net*, 13 juillet 2026.

Lettre ouverte à la Commission européenne, datée du 9 juillet 2026, signée par : Pour Deamin, The Future Society, Future of Life Institute, SaferAI, AI Standards Lab, le Centre pour la Sécurité de l'IA (CeSIA), Apart.

Clement-Fontaine, Mélanie, « L'articulation des stratégies réglementaires de l'IA et des données personnelles : convergence normative ou tension structurelle ? », *Dalloz IP-IT*, 2026.

OCDE, BCG et INSEAD, *The Adoption of Artificial Intelligence in Firms: New Evidence for Policymaking*, OECD Publishing, Paris, 2025.

Y. Bengio, G. Hinton, À. Yao, D. Song, S. Russell, D. Kahneman et al., « Managing extreme AI risks amid rapid progress », *Science* 384, 842–845, 24 mai 2024.

Rapport établi par l'Inspection générale des finances, l'Inspection générale des affaires sociales et l'Inspection générale de l'administration, « Le déploiement de l'intelligence artificielle dans les administrations publiques : analyse comparée, enjeux et conditions de réussite », avril 2026 : proposition n° 4.

Comité interministériel de l'Intelligence artificielle, *Faire de la France une puissance de l'IA*, février 2025 ; À. PALIX & J. BLAIN, « Gouvernance et évolution de la réglementation relative à l'intelligence artificielle : l'approche de la Direction générale des entreprises », *Dalloz IP-IT*, 2026.

CeSIA, « Alignement de l'IA : cartographie des acteurs français », avril 2026.

CeSIA, « Vers une expertise française en évaluation de l'IA : le rôle stratégique de l'INESIA », 2025, p. 3.

G. Stigler, « The Theory of Economic Regulation », *Bell Journal of Economics and Management Science*, 1971. Cité par Clément Duport dans le cadre des travaux de la mission.

Corporate Europe Observatory et LobbyControl, Rapport sur la rédaction du Code de bonnes pratiques sur l'IA à usage général, avril 2025.

AFNOR Normalisation, « Stratégie française de normalisation 2025-2030 : la normalisation, levier d'innovation, de souveraineté et d'influence pour la France », édito de Sébastien Martin, ministre délégué chargé de l'Industrie, p. 3.

« A Blueprint for Multinational Advanced AI Development », note co-rédigée par 29 experts : Abecassis, Adrien, Barry, Jonathan, Bello, Ima, Bengio, Yoshua, Bergeaud, Antonin, Bonnet, Yann, Hacker, Philipp, Harack, Ben, Hatz, Sophia, Henkel, Joachim, Hoos, Holger H., Kitamura, Kit Kit, Lall, Ranjit, Lechelle, Yann, de Leusse, Constance, Martinet, Charles, Miailhe, Nicolas, Morse, Julia C., Negele, Maximilian, Park, Kyung Ryul, Pluckebaum, Miro, Popa-Fabre, Murielle, Prud'homme, Benjamin, Ralle, Yohann, Robinson, Mark, Segerie, Charbel-Raphael, Torreblanca, José-Ignacio, Velasco, Lucia, K. VijayRaghavan, novembre 2025.

OPECST, *Les nouveaux développements de l'intelligence artificielle*, Ass. Nat., n° 642, 29 novembre 2024.

Bernanos, Georges, *La France et les Robots*, Éditions de la France libre, 1946, p. 11.

Lettre de saisine du CCNEN

R E P U B L I Q U E F R A N Ç A I S E

SÉNAT

VANINA
PAOLI-GAGIN

SENATEUR DE L'AUBE

Membre de la commission des finances

Rapporteur des crédits de la Mission Recherche et Enseignement Supérieur

Vice-Président de la Délégation à la Prospective

C.H.E.C.
Promotion Sonia Delaunay

I.H.E.D.N.
S.N.C. 6

M. Claude KIRCHNER
Président
Comité consultatif national d'éthique du numérique (CCNEN)
66 rue de Bellechasse
75007 Paris

Paris, le 11 avril 2026

Monsieur le Président,

Je fais suite à notre réunion du 7 avril au Sénat, en présence de Monsieur Raja Chatila et de Madame Catherine Tessier, pour vous renouveler mes remerciements pour le temps que vous avez bien voulu consacrer et les échanges constructifs que nous avons eus sur l'alignement des systèmes d'intelligence artificielle.

Dans le contexte de cette mission qui m'a été confiée par le Premier ministre, je souhaite par la présente, solliciter l'appui de Comité consultatif national d'éthique du numérique (CCNEN) en vue d'établir un état des lieux, suivi de recommandations, sur l'alignement des systèmes d'IA dans ses trois dimensions (technique, éthique, systémique), au sein des administrations publiques, qu'il s'agisse de l'administration centrale ou des services déconcentrés de l'État.

Dans une perspective d'efficacité et d'efficience, tant humaine, fonctionnelle, que budgétaire, au service de nos administrés, je vous adresse également, par courriel séparé, les éléments qui pourraient éclairer vos travaux en amont sur la nécessaire et préalable réorientation des intentions stratégiques des organisations, ainsi que nous avions pu nous entretenir avec le Ministre délégué David Amiel, le 5 février dernier.

Au regard du délai assez court donc nous disposons dans le cadre de la mission pour remettre notre rapport final, il nous serait utile que cet avis du CCNEN puisse nous être retourné avant le vendredi 12 juin.

15, rue de Vaugirard - 75291 PARIS CEDEX 06 - Telephone : 01 42 34 40 38 - v.paoli-gagin@senat.fr – Twitter/X :@GaginPaoli

Dans l'attente de votre réponse, et me tenant à la disposition du Comité pour toute question afférant à cette saisine, je vous prie de bien vouloir agréer, Monsieur le Président, l'expression de ma considération distinguée.

.

Vanina PAOLI-GAGIN
Sénateur de l'Aube

15, rue de Vaugirard - 75291 PARIS CEDEX 06 - Telephone : 01 42 34 40 38 - v.paoli-gagin@senat.fr – Twitter/X :@GaginPaoli

Étude comparative internationale

Liste des 12 pays :

Canada	États-Unis	Brésil	Allemagne
Royaume-Uni	Maroc	Émirats arabes unis	Israël
Inde	Chine	Singapour	Corée du Sud

Cette liste a été établie en concertation avec Madame la Ministre Anne Le Hénanff et Madame l'Ambassadrice Clara Chappaz.

Le choix de ce panel répond à un objectif de couverture géographique large, afin d'intégrer des approches issues de l'Amérique du Nord, de l'Asie, du Moyen-Orient, de l'Amérique latine et de l'Afrique. Il vise également à documenter des contextes nationaux différenciés en matière de stratégie IA, de régulation, de capacités industrielles, de positionnement international et de rapport entre puissance publique et acteurs privés.

Questionnaire de l'étude

Question 1 – Qualification du sujet dans le débat public et institutionnel

Comment la notion d'alignement des systèmes d'intelligence artificielle est-elle appréhendée dans le pays concerné ?

Merci de préciser, dans la mesure du possible :

- si cette notion fait l'objet d'un usage explicite dans le débat public, institutionnel, scientifique ou économique ;
- les termes éventuellement utilisés localement pour désigner des problématiques proches (sécurité de l'IA, robustesse, fiabilité, contrôlabilité, gouvernance, évaluation, etc.) ;
- le niveau de priorité accordé à ce sujet par les autorités publiques et les principales parties prenantes ;
- les secteurs ou usages dans lesquels ce sujet est considéré comme particulièrement sensible ou stratégique.

Question 2 – Doctrine publique, stratégie nationale et gouvernance

Existe-t-il dans le pays concerné une doctrine publique, une stratégie nationale ou un cadre de gouvernance spécifique relatif à l'alignement, à la sécurité ou à la maîtrise des systèmes d'IA avancés ?

Merci de préciser, dans la mesure du possible :

- les orientations stratégiques adoptées par les autorités publiques ;
- les administrations, agences, autorités ou instances en charge du sujet ;

- le rôle respectif de l'État, des régulateurs, du secteur privé et du monde académique ;
- l'existence éventuelle de comités d'experts, groupes de travail ou dispositifs de coordination dédiés.

Question 3 – Articulation avec le cadre réglementaire existant

Comment les enjeux liés à l'alignement des systèmes d'IA s'articulent-ils avec les cadres réglementaires et normatifs existants dans le pays concerné ?

Merci de préciser, dans la mesure du possible :

- les principaux textes, normes ou lignes directrices mobilisés ;
- l'articulation entre les enjeux d'alignement et les obligations relatives à la protection des données, à la cybersécurité, à la sécurité des produits, à la responsabilité ou à la transparence ;
- l'existence éventuelle de débats sur le risque de sur-réglementation ou, à l'inverse, sur l'insuffisance du cadre existant ;
- les évolutions envisagées à court ou moyen terme.

Question 4 – Écosystème d'acteurs et état de l'art

Quels sont, dans le pays concerné, les principaux acteurs impliqués dans la recherche, le développement, l'évaluation, le déploiement ou la gouvernance de solutions contribuant à l'alignement des systèmes d'IA ?

Merci de préciser, dans la mesure du possible :

- les principaux laboratoires de recherche, universités ou centres d'expertise ;
- les entreprises concernées (grands groupes, PME, start-ups) ;
- les organismes d'évaluation, d'audit, de normalisation ou de certification ;
- les organisations professionnelles, fondations, ONG ou autres structures influentes sur le sujet.

Question 5 – Méthodes d'évaluation, audit, certification et supervision

Quelles sont les approches concrètes d'évaluation, d'audit, de certification ou de supervision des systèmes d'IA mises en œuvre, envisagées ou débattues dans le pays concerné ?

Merci de préciser, dans la mesure du possible :

- les méthodes ou outils utilisés pour évaluer la robustesse, la sécurité, la fiabilité, l'explicabilité ou la contrôlabilité des systèmes ;
- l'existence de référentiels, standards, protocoles de test, benchmarks ou labels ;
- les acteurs chargés de ces fonctions d'évaluation ou de supervision ;
- les limites identifiées, notamment en matière de faisabilité technique, de coût, de modèle économique ou de périmètre d'intervention.

Question 6 – Structuration de la demande et dynamique de marché

Observe-t-on dans le pays concerné l'émergence d'une demande identifiable pour des systèmes d'IA plus sûrs, plus fiables, plus auditables ou plus alignés ?

Merci de préciser, dans la mesure du possible :

- quels acteurs expriment cette demande (administrations, grands donneurs d'ordre, secteurs régulés, investisseurs, consommateurs, etc.) ;
- dans quels secteurs ou cas d'usage cette demande se manifeste le plus fortement ;
- si cette demande se traduit par des exigences contractuelles, des dispositifs d'achat, des référentiels ou des critères de sélection ;
- les principaux freins à la structuration d'un marché sur ces enjeux.

Question 7 – Leviers publics de soutien, financement et politique industrielle

Quels sont les leviers d'action publique mobilisés dans le pays concerné pour soutenir les capacités nationales en matière d'IA sûre, fiable, contrôlable ou alignée ?

Merci de préciser, dans la mesure du possible :

- les dispositifs de financement de la recherche et de l'innovation ;
- les programmes publics ou partenariats public-privé existants ;
- le rôle éventuel de la commande publique ;
- les initiatives visant à structurer une chaîne de valeur nationale ou à renforcer l'autonomie stratégique du pays sur ces sujets.

Question 8 – Enseignements, perspectives et appréciation d'ensemble

Quels sont, selon les interlocuteurs locaux, les principaux enseignements, facteurs de succès, points de vigilance ou perspectives d'évolution en matière d'alignement, de sécurité ou de gouvernance des systèmes d'IA dans le pays concerné ?

Merci de préciser, dans la mesure du possible :

- les tendances émergentes ;
- les débats structurants ou controverses éventuelles ;
- les priorités identifiées pour les prochaines années ;
- les éléments pouvant présenter un intérêt particulier dans une perspective française ou européenne.

Commentaires éventuels du pays questionné :

.....
.....
.....
.....

Scenarii

L'inertie enchaînée

La France et l'Europe appliquent les cadres existants sans stratégie spécifique d'alignement.

Ce scénario prolonge les constats de fragmentation relevés dans les auditions. Les acteurs distinguent mal alignement, IA de confiance, conformité, sécurité, souveraineté et évaluation. L'INESIA est identifié comme utile mais encore fragile, sans structure juridique forte, sans moyens humains dédiés suffisants et avec une coordination encore limitée. Le RIA fixe des obligations, mais ne suffit pas à lui seul à produire des méthodes robustes, vérifiables et opérationnelles.

Risques : fragmentation persistante ; dépendance accrue aux clouds, modèles, GPU et couches logicielles extra-européennes ; impossibilité d'évaluer les modèles de frontière et les systèmes agentiques ; affaiblissement de la confiance des entreprises, administrations et citoyens ; difficulté à faire émerger un marché européen de l'IA alignée.

Le rattrapage performatif

La France décide d'investir dans l'évaluation, la vérification, la recherche et la gouvernance, sur le modèle centralisé du UK AISI.

Ce scénario repose sur un programme national de recherche sur le *safe by design*, une plateforme d'évaluation mutualisée, le renforcement de l'INESIA ou d'une instance de coordination dotée de moyens, d'ETP, d'un mandat clair et d'un accès aux modèles à tester. Il implique également un soutien aux acteurs industriels, startups, laboratoires et organismes d'évaluation, ainsi que la structuration d'un marché de l'IA alignée fondé sur la preuve, l'auditabilité, la conformité, la cybersécurité, l'explicabilité et la responsabilité.

Risques : retard méthodologique si les capacités restent dispersées ; risque de dépendre des évaluations privées des fournisseurs de modèles ; faible adoption par les entreprises si l'alignement reste un vocabulaire d'experts ; risque de multiplication des labels sans référentiel commun.

Le repli résilient

La France accepte que des chocs puissent survenir et prépare des capacités d'amortissement en mode dégradé, avec placement de l'humain au centre.

Les modèles changent, les usages se déplacent et les capacités agentiques accroissent l'incertitude. Ce scénario repose sur la nécessité d'une résilience sociétale, au-delà de la seule régulation des fournisseurs.

Risques : Choc cyber ou attaque adverse sur un système IA ; rupture d'accès à un fournisseur, un cloud, un modèle ou une puce ; exfiltration de données sensibles, notamment code source et données industrielles ; perte de compétences humaines par automatisation excessive et perte de compétitivité.

La puissance alignée

La France et l'Europe font de l'alignement une politique industrielle, scientifique et stratégique.

L'alignement devient un levier de puissance et d'industrialisation. La France et l'Europe produisent des méthodes, outils, standards, auditeurs et référentiels susceptibles d'acquérir un effet extraterritorial. La maîtrise de la chaîne de valeur s'appuie sur la commande publique et la BITN-IA.

Risques : souveraineté déclarative ou *sovereignty-washing* si la chaîne de valeur n'est pas maîtrisée ; risque de fragmentation européenne si les États membres ne partagent pas la même définition de la souveraineté et de l'alignement ; acceptabilité sociale des projets industriels, notamment les data centers ; manque de moyens financiers.