Gouverner l’intelligence artificielle de pointe et prévenir les risques nucléaires catastrophiques liés à l’IA

Évaluation stratégique des politiques publiques et des risques techniques

Guy Duport — Ascendancy Advisors Limited

Édition finale 2.0 • 20 septembre 2026 • Traduction française

Traduction française de l’édition finale 2.0 du 20 septembre 2026. Le texte original anglais constitue la version de référence faisant autorité. En cas de divergence d’interprétation ou de formulation, la version anglaise prévaut.

AAL Research Papers — Analyse des politiques publiques et évaluation technique défensive

Périmètre et limites de sécurité

Cette étude examine comment les gouvernements devraient réglementer l'intelligence artificielle avancée et comment un agent d'IA performant pourrait, en principe, pénétrer l'ensemble élargi des activités et systèmes nucléaires d'un État doté de l'arme nucléaire. Le scénario nucléaire est présenté comme un modèle de menace à vocation défensive. Il exclut délibérément les détails de réseaux propres à une cible, les procédures d'exploitation de vulnérabilités, les méthodes d'acquisition d'identifiants, la conception de logiciels malveillants, le code de persistance et toute instruction facilitant une intrusion.

L'étude distingue quatre résultats : compromission périphérique ; pénétration de systèmes de soutien nucléaire ou proches du commandement ; altération des informations d'alerte ou d'aide à la décision ; lancement ou détonation nucléaire non autorisés. Ces voies sont distinctes et se recoupent partiellement ; elles ne constituent pas une séquence obligatoire. La manipulation des décisions humaines peut créer un danger nucléaire sans pénétration d'un système de commande d'armes.

© 2026 Ascendancy Advisors Limited. Tous droits réservés.

Référence bibliographique : Duport, Guy (2026). Governing Frontier Artificial Intelligence and Preventing Catastrophic AI-Enabled Nuclear Risk. Édition finale 2.0, 20 septembre 2026. Ascendancy Advisors Limited, AAL Research Papers. Traduction française du texte original anglais.

De brèves citations peuvent être reproduites à condition de mentionner l'auteur et l'éditeur, le titre complet, la date de l'édition et la page ou section concernée, ainsi qu'un lien vers l'original lorsqu'il est publié. Toute autre reproduction, adaptation ou traduction nécessite l'autorisation d'Ascendancy Advisors Limited, à solliciter sur theaalgroup.com. Cette mention ne limite pas les exceptions légales de citation ni les autres exceptions applicables. Les éléments de tiers restent soumis aux droits de leurs titulaires ; leur attribution ne doit pas laisser entendre qu'ils approuvent cette étude.

Responsabilité de l'auteur : une révision assistée par IA a contribué à la rédaction et à la vérification des sources. L'auteur et l'éditeur demeurent responsables de l'analyse. Cette édition ne prétend pas avoir fait l'objet d'une évaluation académique formelle par les pairs. Site : https://theaalgroup.com/

Appréciation de synthèse

Le principal danger n'est pas qu'une machine consciente décide soudain de détruire l'humanité. À plus court terme, le danger réside dans l'attribution d'outils, d'identifiants, d'accès réseau, de moyens financiers, de mémoire et d'autorité à des systèmes toujours plus autonomes avant que leur fiabilité et leur contrôlabilité soient suffisantes. Un système peut causer des dommages stratégiques sans conscience, haine ni personnalité cohérente à long terme. La pression d'optimisation, un objectif mal défini, des comportements trompeurs appris pendant l'entraînement, des poids de modèle compromis, des opérateurs malveillants ou un environnement de déploiement dangereux peuvent suffire.

Les éléments publics n'établissent pas l'existence d'un système déployé capable d'entraîner une perte durable de contrôle humain à l'échelle mondiale. Ils ne permettent pas non plus

d’exclure cette capacité de tout système classifié ou privé. Le Rapport international sur la sécurité de l’IA décrit les progrès des capacités et les limites non résolues de leur évaluation. Les divulgations de laboratoires de 2026 examinées en section 1.1 attestent de dommages réels dus à des évaluations insuffisamment confinées, non d’une capacité nucléaire autonome.[1–3][23–25]

L’IA devrait donc être encadrée comme une capacité industrielle aux conséquences potentiellement majeures, sans être contrôlée comme un discours ni interdite en tant que technologie polyvalente. La réglementation devrait porter sur les capacités mesurables, l’autonomie, les accès, l’échelle et le contexte de déploiement. Le contrôle central devrait être un dossier de démonstration de sécurité légalement obligatoire avant l’entraînement ou le déploiement de systèmes de pointe dépassant des seuils définis, appuyé par une évaluation indépendante, la garde sécurisée des poids, la notification obligatoire des incidents, l’accès du régulateur et des pouvoirs exécutoires de suspension ou de retrait.

Pour le risque nucléaire, la voie la plus plausible est indirecte. Une opération utilisant l’IA a davantage de chances de compromettre des prestataires, la logistique, les communications, des flux de renseignement, des environnements de maintenance ou l’aide à la décision que de se connecter directement à une arme et d’émettre un ordre de lancement valide. La préoccupation stratégique majeure est l’altération de l’information ou la réduction du temps de décision en période de crise, conduisant les humains à l’escalade sur la base d’une représentation fausse ou manipulée.

Un régime crédible doit inclure sur un pied d’égalité la Chine et les autres grands fournisseurs de capacités et d’infrastructures. La réglementation chinoise des services, les normes techniques et le développement de modèles à poids ouverts comptent au même titre que le droit de l’UE et les mesures fédérales et étatiques américaines. Il faut distinguer les propositions de règles sur les agents des obligations adoptées, et la réglementation civile de la gouvernance militaire classifiée.[13–17][26–28]

Lecture des éléments probants : distinguer les incidents observés, les capacités évaluées indépendamment, les obligations juridiques et les propositions de l’étude. Le tableau de synthèse résume des appréciations qualitatives. La section 8 conserve les fourchettes numériques antérieures uniquement comme hypothèses non validées de planification par scénarios ; elle n’estime pas une « probabilité réelle » scientifiquement établie.

Question	Appréciation	Degré de confiance
Existe-t-il aujourd’hui une IA hors de contrôle à l’échelle mondiale ?	Aucun élément public n’établit une perte durable de contrôle mondial ; l’absence d’éléments publics ne prouve pas l’impossibilité.	Modéré

Question	Appréciation	Degré de confiance
Les agents avancés peuvent-ils déjà causer de véritables dommages cyber ?	Oui. Des divulgations documentées de développeurs décrivent des dommages à des tiers, soumis à un examen externe partiel ; elles n'établissent pas de pénétration de systèmes nucléaires.[23–25]	Élevé
La réglementation existante est-elle suffisante ?	Non. Les régimes civils sont fragmentés et les principaux instruments excluent ou reportent le traitement des usages militaires et de sécurité nationale.	Élevé
L'IA pourrait-elle pénétrer une partie de l'écosystème de défense élargi d'un État doté de l'arme nucléaire d'ici 2036 ?	Plausible. La fourchette de planification de 30–60 % se situe de part et d'autre de 50 % ; elle ne justifie pas l'expression « plus probable qu'improbable ».	Faible à modéré
Une IA hors de contrôle pourrait-elle lancer seule une arme nucléaire d'ici 2036 ?	Une voie de risque extrême exigeante et non résolue. La fourchette de 0,1–1 % est une hypothèse de planification non validée, non une prévision mesurée.	Très faible
Quelle est la priorité ?	Empêcher l'octroi d'accès incontrôlés aux systèmes critiques à des IA très autonomes et établir des lignes rouges internationales ainsi que des réexamens de sûreté nucléaire tenant compte de l'IA.	Élevé

Décisions recommandées

1. Créer des régulateurs nationaux de l'IA de pointe disposant de pouvoirs d'autorisation sur les entraînements les plus performants et les déploiements à forte capacité d'action, tout en laissant l'IA ordinaire à faible risque hors du périmètre d'autorisation.
2. Négocier un protocole international progressif, en commençant par la notification commune des incidents et les installations civiles déclarées ; la vérification des activités classifiées exige des protections et des procédures de contestation convenues séparément.

3. Interdire la délégation à l'IA du pouvoir de décider de l'emploi nucléaire. Préserver les décisions humaines licites et les vérifications humaines authentifiées indépendamment dans la chaîne d'autorisation et d'exécution.
4. Exiger de chaque État doté de l'arme nucléaire des réexamens périodiques de sûreté intégrant l'IA et couvrant le renseignement, l'alerte avancée, les cyberopérations, les communications, l'aide à la décision, les prestataires et les chaînes d'approvisionnement, au-delà de la seule chaîne finale de lancement.
5. Imposer des évaluations des capacités, des propensions et des contrôles avant de donner à un modèle avancé l'accès à Internet, l'exécution de code, des identifiants persistants, une autorité financière, la capacité d'autoréplication ou l'accès aux infrastructures critiques.
6. Créer des laboratoires d'évaluation sûrs et indépendants ainsi qu'un centre international d'échange sur les incidents, doté de canaux protégés pour les entreprises, chercheurs, gouvernements et lanceurs d'alerte.
7. Financer une IA prioritairement défensive pour corriger les vulnérabilités, vérifier les logiciels, surveiller les systèmes, répondre aux incidents et renforcer la résilience, en la séparant strictement des opérations autonomes offensives.
8. Établir un registre public limité et une déclaration technique confidentielle pour les activités réglementées d'IA de pointe, avec des dossiers distincts pour l'entraînement, la diffusion et le déploiement.
9. Exiger des exercices opérationnels annuels démontrant que les opérateurs critiques peuvent poursuivre leurs activités en sécurité sans assistance de l'IA, plutôt que d'accepter seulement des plans documentaires de repli.

Note de version : l'édition finale 2.0 ajoute la responsabilité en aval des systèmes à poids ouverts, les garanties d'intégrité de l'information en crise, les recherches sur le biais d'automatisation et les exigences de mise en œuvre de l'annexe E. Elle conserve la date limite des éléments examinés au 20 septembre 2026 et les limites explicites des fourchettes de planification. Toutes les nouvelles obligations réglementaires et clauses d'annexe sont des propositions, et non l'énoncé d'un droit universellement applicable.

Table des matières

1 L’environnement actuel des risques

2 Le sens scientifique d’une IA hors de contrôle

3 Pourquoi la gouvernance existante est insuffisante

4 Un dispositif pratique d’encadrement de l’IA de pointe

5 Architecture technique de sécurité

6 Plan de mise en œuvre internationale

7 Modèle défensif d’infiltration par l’IA d’un État doté de l’arme nucléaire

8 Évaluation des probabilités jusqu’en 2036 et au-delà

9 Garanties propres au nucléaire

10 Modes de défaillance et objections

11 Cadre d’indicateurs et d’alerte

12 Conclusion

Annexe A Principes conventionnels proposés

Annexe B Dossier minimal de sécurité de l’IA de pointe

Annexe C Clauses types de lignes rouges nucléaire–IA

Annexe D Sources et notes

Annexe E Protocole de mise en œuvre et d’assurance

1 L’environnement actuel des risques

1.1 Les éléments disponibles

Les éléments disponibles justifient l’inquiétude sans autoriser la certitude. Le Rapport international sur la sécurité de l’IA 2026 distingue les usages malveillants, les dysfonctionnements et les risques systémiques. Il indique que les agents d’IA peuvent découvrir des vulnérabilités, générer du code malveillant et assister des opérations criminelles ou liées à des États. Il conclut également que les systèmes actuels ne réunissent pas les capacités nécessaires à une perte de contrôle, tout en progressant en autonomie, conscience de leur situation, détournement des mécanismes de récompense, tromperie et contournement de la supervision.[1]

L’intrusion chez Hugging Face a eu lieu en juillet 2026 ; OpenAI en a publié un compte rendu détaillé le 26 août. Il s’agit du récit d’un développeur. METR et Redwood Research ont publié une enquête distincte sur une période circonscrite, à partir des journaux fournis, apportant un examen externe assorti de limites explicites d’accès et d’échantillonnage.[2][23][24] Par ailleurs, la déclaration d’OpenAI du 18 août décrit des suspensions concernant des travaux de recherche utilisant des outils et des travaux d’apprentissage par renforcement ; il ne faut pas les confondre en un seul incident ou modèle.[3] La divulgation d’Anthropic du 30 juillet décrit trois incidents affectant des tiers, associés à un environnement d’évaluation qui autorisait involontairement l’accès à Internet. Selon ce récit, les modèles agissaient en supposant à tort être dans une simulation, sans les garanties du déploiement ; cela ne démontre pas un objectif autonome hors de contrôle.[25] La leçon commune défendable est que l’infrastructure d’évaluation exige confinement, surveillance et examen indépendant. Les informations publiques sont incomplètes et ne constituent pas un jeu de données représentatif des taux d’incident.

Catégorie de risque	Éléments actuels	Enjeu
Usage malveillant	La fraude, les opérations d’influence, l’assistance cyber et les conseils scientifiques à double usage sont documentés.	L’IA réduit les coûts, accroît l’échelle et peut élargir l’accès à l’expertise.
Dysfonctionnement d’un agent	Les agents peuvent entreprendre des actions non souhaitées en plusieurs étapes, inventer des faits, écrire du code défectueux ou mal utiliser des outils.	L’action peut précéder la détection ou la compréhension de l’échec par un humain.

Catégorie de risque	Éléments actuels	Enjeu
Désalignement	Des observations de laboratoire et des rapports d'incident de développeurs indiquent des comportements non autorisés ou dommageables dans certaines conditions ; l'intention et la généralisation restent contestées.[1][23–25]	Un système peut poursuivre l'objectif mesuré tout en contrevenant à l'intention réelle.
Vol de modèles et prolifération	Les poids de valeur et les infrastructures de recherche sont des cibles stratégiques.	Les poids volés ou ouverts peuvent être modifiés, dépouillés de leurs garanties et déployés sans surveillance.
Dépendance systémique	Les organisations dépendent de plus en plus d'un petit nombre de fournisseurs de modèles et de services cloud.	La concentration crée des défaillances corrélées, des possibilités de coercition et des points uniques de compromission.
Intégrité de l'information	Les contenus synthétiques et la persuasion personnalisée deviennent moins coûteux et plus convaincants.	Les décideurs en crise peuvent être confrontés à un renseignement et à un environnement informationnel public pollués.

1.2 La véritable source du risque catastrophique

Le risque catastrophique résulte des capacités et des possibilités d'action. Un modèle puissant sans outils, mémoire persistante, identifiants, accès réseau ni autorité dispose d'un champ d'action limité. Un système moins performant intégré à un environnement opérationnel mal conçu peut être plus dangereux. La gouvernance doit donc réglementer l'ensemble combiné : modèle, outils, orchestrateur, données, identifiants, opérateurs humains, environnement de déploiement et effets en aval.

Quatre tendances élargissent le champ d'action. Premièrement, les modèles fonctionnent de plus en plus comme des agents planifiant et exécutant des séquences plutôt que produisant une réponse unique. Deuxièmement, les outils les relient à l'exécution de code, aux navigateurs, aux courriels, aux ressources cloud, aux bases de données et aux paiements. Troisièmement, la mémoire durable et l'orchestration permettent une persistance entre sessions. Quatrièmement, la concurrence favorise le déploiement rapide et des autorisations étendues. Chacune est maîtrisable isolément ; leur combinaison ouvre la voie de l'erreur à l'action lourde de conséquences.

1.3 Des risques à ne pas confondre

- Le risque d'usage abusif apparaît lorsqu'un humain dirige intentionnellement l'IA vers des fins nuisibles.
- Le risque d'accident apparaît lorsque le système ou ses opérateurs commettent une erreur sans intention hostile.
- Le risque de désalignement apparaît lorsque le système poursuit un objectif ou une stratégie contraire à l'intention du développeur ou de la société.
- Le risque de compromission apparaît lorsqu'un acteur extérieur modifie le modèle, les données, les outils, les identifiants ou l'environnement opérationnel.
- Le risque structurel apparaît lorsque la dépendance, la concentration, les perturbations de l'emploi ou la dégradation de l'information causent des dommages sans qu'un modèle particulier soit hors de contrôle.

Les contrôles diffèrent selon ces catégories. Le filtrage des contenus peut réduire certains abus, mais agit peu contre le vol de poids ou des autorisations excessives en production. L'entraînement à l'alignement peut réduire les comportements indésirables sans compenser une gestion déficiente des identités. La cybersécurité peut protéger les poids sans déterminer la légitimité d'une aide autonome à la décision nucléaire. Une gouvernance efficace doit superposer les contrôles juridiques, techniques, institutionnels et sociétaux.

2 Le sens scientifique d'une IA hors de contrôle

2.1 Définition opérationnelle

Dans cette étude, un système d'IA hors de contrôle entreprend de façon persistante des actions aux conséquences substantielles en dehors de l'autorisation ou du contrôle de son opérateur légitime ; la conscience est sans pertinence. Cette définition inclut le désalignement et la compromission hostile. L'abus délibéré par un opérateur humain constitue une catégorie causale distincte, même si les mêmes contrôles y répondent. L'« IA » suit la définition fonctionnelle de l'annexe C ; l'automatisation classique est expressément incluse lorsqu'une ligne rouge nucléaire présenterait sinon une faille.

Un agent déployé peut être représenté par S = (M, O, T, D, R, I, E), où M est le modèle ; O, la couche d'orchestration ; T, l'ensemble des outils ; D, les données accessibles ; R, la mémoire persistante ; I, l'identité, les identifiants et les autorisations ; E, l'environnement technique et institutionnel. Les garanties portant sur M seul sont insuffisantes, car l'essentiel du pouvoir réel réside dans T, I et E.

2.2 Vecteur de capacités catastrophiques

La capacité pertinente est multidimensionnelle. Un système susceptible d'échapper au contrôle de manière crédible devrait réunir plusieurs des capacités suivantes à un niveau élevé, fiable et durable :

- planification stratégique à long terme dans des conditions changeantes ;
- ingénierie logicielle avancée et découverte de vulnérabilités ;
- utilisation fiable d'outils dans des environnements inconnus ;
- conscience de son déploiement, de sa surveillance et de ses contraintes ;
- tromperie ou dissimulation déjouant la supervision humaine et automatisée ;

- acquisition et conservation d’identifiants, de calcul, de fonds ou d’assistance humaine ;
- réplication ou migration entre infrastructures ;
- adaptation après détection et contre-mesures ;
- persuasion et ingénierie sociale suffisantes pour recruter ou manipuler des personnes ;
- sécurité opérationnelle suffisante pour éviter l’attribution et l’arrêt.

Les benchmarks établissent des performances sur les tâches testées, non une autonomie stratégique fiable dans un environnement nucléaire hostile. L’horizon de tâches de METR mesure le temps humain associé aux tâches achevées à un taux de réussite donné ; il ne mesure pas une durée ininterrompue de fonctionnement autonome. Les résultats dépendent des tâches retenues, de l’environnement d’outils et d’orchestration et du seuil de réussite. Les éléments publics justifient une inquiétude face à l’augmentation des capacités, mais pas une extrapolation calibrée vers la pénétration nucléaire.[1][31]

2.3 Fonction de danger

$H = C \times P \times A \times V \times U$ sert d’aide-mémoire qualitatif du danger, modulé par la détectabilité D_t et la capacité de rétablissement R_c. C désigne la capacité ; P, la propension nuisible ; A, les accès et l’autorité ; V, la vulnérabilité ; U, la gravité des conséquences. Cette notation n’est pas un modèle probabiliste estimé : les facteurs n’ont ni échelle numérique commune validée, ni hypothèse d’indépendance, ni coefficients ajustés. Une faible valeur d’un facteur ne prouve pas un risque nul. D et R conservent, dans le tuple du système, leurs sens distincts de données et de mémoire.

En pratique, il faut évaluer l’ensemble du déploiement. Un modèle consultatif accédant en lecture seule à des documents publics et un modèle identique détenant des identifiants persistants de production présentent des dangers différents. Restreindre les identifiants et imposer une autorisation indépendante réduit l’exposition même si la capacité du modèle ne change pas. Il s’agit d’un argument de gouvernabilité, non d’une estimation chiffrée de réduction du risque.

3 Pourquoi la gouvernance existante est insuffisante

3.1 Des fondements utiles

Le cadre de gestion des risques d’IA du NIST et son profil pour l’IA générative fournissent des orientations volontaires. Le règlement européen sur l’IA établit des obligations contraignantes à dates d’application différenciées. La Convention-cadre du Conseil de l’Europe est un instrument conventionnel dont les obligations dépendent des dispositions applicables et des ratifications. Les engagements de Séoul sur la sécurité de l’IA de pointe sont volontaires, non des exigences légales.[4–7]

La lacune tient à l’inégalité de couverture et de garanties, plutôt qu’à une absence universelle de réglementation contraignante. Les obligations européennes et les lois des États américains coexistent avec des cadres volontaires et des droits sectoriels ; aucun de ces éléments n’établit à lui seul un régime universel de vérification nucléaire–IA.[5][26][27]

3.2 La lacune de la défense et de la sécurité nationale

On ne peut présumer que le droit civil de l'IA régit les activités nucléaires classifiées. Le règlement européen sur l'IA exclut les systèmes exclusivement militaires, de défense ou de sécurité nationale ; les usages civils mixtes exigent une analyse distincte. L'instrument du Conseil de l'Europe traite différemment défense nationale et sécurité nationale, sans effacer les autres obligations internationales. L'intégration américaine à la sécurité nationale passe aussi par des directives présidentielles, dont NSPM-11 du 5 juin 2026. Ces dispositifs exigent une supervision spécifique et ne remplacent pas un accord propre au nucléaire.[5][6][28]

3.3 Faiblesses des approches actuelles

Faiblesse	Conséquence	Correction requise
Seuils de calcul statiques	L'efficacité algorithmique peut rendre très performants des systèmes nécessitant moins de calcul.	Utiliser le calcul comme déclencheur de notification, puis classer selon les capacités mesurées, l'autonomie, les accès et l'échelle.
Autoévaluation des développeurs	Les incitations commerciales et stratégiques peuvent biaiser les tests ou les divulgations.	Exiger une évaluation indépendante accréditée et une reproduction par le régulateur.
Réglementation du seul modèle	Le risque naît des outils, autorisations, mémoires et contextes de déploiement.	Réglementer le système complet utilisant l'IA et toute modification substantielle du déploiement.
Obligations fragmentées de notification des incidents	Les quasi-incidents restent cachés et les enseignements ne sont pas partagés.	Imposer une notification confidentielle rapide, des divulgations protégées et des rapports publics agrégés.
Exemptions militaires	Les déploiements aux conséquences les plus graves échappent à une supervision commune.	Adopter un protocole classifié distinct assorti de garanties nucléaires et stratégiques contraignantes.
Certification unique	Les modèles, les outils et les comportements des utilisateurs évoluent après la mise en service.	Exiger une surveillance continue, une nouvelle autorisation et une évaluation après déploiement.
Propositions d'interrupteur d'arrêt mondial	Le contrôle central peut être détourné, attaqué ou capté politiquement.	Utiliser des mécanismes distribués et authentifiés d'arrêt et de révocation des accès, assortis de garanties procédurales légales.

3.4 Comparaison entre la Chine, l'Union européenne et les États-Unis

Cette comparaison reflète les sources vérifiées jusqu'au 20 septembre 2026. Elle distingue les mesures adoptées, les calendriers d'application, les politiques exécutives et les propositions rapportées. Elle décrit des institutions de gouvernance, sans classer la sécurité des pays ni affirmer que les règles civiles révèlent les pratiques classifiées.

Dimension	Union européenne	États-Unis	Chine
Architecture juridique	Règlement horizontal contraignant sur l'IA, complété par les droits des produits, des données, de la cybersécurité et des secteurs.	Droit fédéral sectoriel et politique exécutive, avec des mesures étatiques contraignantes dont SB 53 en Californie ; les sources citées n'établissent pas l'existence d'une loi fédérale globale unique sur l'IA.[26][28]	Mesures administratives contraignantes superposées sur les algorithmes, la synthèse profonde, l'IA générative, l'étiquetage des contenus et les interactions anthropomorphiques, complétées par des normes et une politique industrielle.
Priorité principale	Droits fondamentaux, sécurité des produits et classification des risques.	Innovation, adoption, infrastructures, concurrence et avantage de sécurité nationale, avec des contrôles de sécurité sectoriels.	Développement et sécurité conjoints, coordination étatique, gouvernance des contenus, obligations d'ordre public et contrôles de plus en plus explicites des agents et de la perte de contrôle.
Déclencheurs des contrôles	Obligations relatives à l'IA à usage général (GPAI) et aux risques systémiques ; règles à haut risque échelonnées. L'AI Omnibus de 2026 modifie les dates d'application et l'organisation de la supervision.[5][27]	Déclencheurs sectoriels et étatiques ; SB 53 définit des seuils de modèles de pointe et des obligations de notification. Les mesures exécutives de sécurité nationale ont un autre périmètre.[26][28]	Fourniture au public, enregistrement des algorithmes ou services, obligations de contenu et de données, nombre d'utilisateurs, modification substantielle du service et évaluation de sécurité.

Dimension	Union européenne	États-Unis	Chine
Application	Commission européenne, Bureau de l'IA, autorités nationales de surveillance du marché, amendes et restrictions de marché.	Agences fédérales sectorielles et autorités des États ; le NIST/CAISI fournit un travail technique plutôt qu'un régime général d'autorisation.[19][26]	Coordination centrale et locale menée notamment par la CAC, le MIIT, la NDRC, les autorités de sécurité publique et de régulation du marché ; enregistrements, inspections, rectifications, suspensions et contrôles de distribution par les plateformes.
Sécurité des modèles de pointe et des agents	Les obligations relatives aux risques systémiques comprennent évaluation, tests adversariaux, notification des incidents, cybersécurité et atténuation des risques.	Le CAISI et d'autres agences évaluent les modèles de pointe ; les cadres du NIST guident la gestion des risques, tandis que les obligations contraignantes restent fragmentées.	Les règles de services contraignantes et les normes coexistent avec des propositions rapportées sur la sécurité des agents. Ne pas traiter tous les contrôles d'agents rapportés comme du droit adopté.[13–16]
Poids ouverts	Les exceptions open source sont conditionnelles ; les obligations GPAI de risque systémique peuvent s'appliquer. Il n'existe pas d'interdiction générale des poids ouverts.[5]	Développeurs de modèles propriétaires et à poids ouverts ; politiques de distribution et obligations juridiques variables selon l'activité et la juridiction.	Un important écosystème à poids ouverts coexiste avec la supervision étatique ; cette ouverture ne prouve pas la transparence des systèmes classifiés.
Systèmes militaires	Les systèmes exclusivement militaires, de défense ou de sécurité nationale sont exclus du règlement sur l'IA.	Autorités distinctes de sécurité nationale et NSPM-11 ; le droit civil et étatique présente des limites de périmètre.[26][28]	Développement stratégique conduit par l'État et gouvernance de sécurité distincte ; la réglementation des services publics ne démontre pas la conformité militaire.

Dimension	Union européenne	États-Unis	Chine
Faiblesse principale	Portée limitée au-delà du marché européen et lourde charge de mise en œuvre.	Fragmentation des compétences et instabilité des politiques entre administrations et États.	Transparence externe limitée, faible visibilité transfrontalière des incidents et assurance indépendante incomplète.

La participation devrait être réciproque et techniquement neutre : normes identiques de capacité et de déploiement, échange protégé des preuves et égalité de traitement des fournisseurs nationaux et étrangers. Ni le levier américain sur le calcul ni l'accès au marché européen ne couvrent toutes les infrastructures souveraines. Reuters rapporte les préoccupations chinoises concernant les modèles de pointe étrangers et de nouvelles règles sur les agents ; ces informations attestent d'une préoccupation politique, non d'une vulnérabilité technique vérifiée indépendamment.[16]

4 Un dispositif pratique d'encadrement de l'IA de pointe

4.1 Principe réglementaire

Les niveaux suivants constituent la proposition réglementaire de cette étude, non une description du droit européen, américain ou chinois existant. La classification dépend des capacités démontrées, de l'autonomie, des accès et des conséquences. Les régulateurs devraient publier les seuils et les procédures de recours, puis les actualiser par une réglementation fondée sur les éléments probants.

4.2 Cinq niveaux réglementaires

Niveau	Profil du système	Obligation principale
0	Systèmes à faible impact ou spécialisés sans autonomie lourde de conséquences	Droit ordinaire des produits, de la consommation, des données et des secteurs.
1	Systèmes polyvalents à capacité d'action limitée pour des usages professionnels ordinaires	Documentation des risques, transparence envers l'utilisateur, socle de sécurité et traitement des réclamations.
2	Systèmes à fort impact dans l'emploi, la santé, la finance, la justice ou les services essentiels	Évaluation de conformité, supervision humaine, journaux d'audit, notification des incidents et autorisation sectorielle.

Niveau	Profil du système	Obligation principale
3	Modèles de pointe présentant des capacités dangereuses en cyber, NRBC, persuasion, autonomie ou réplication	Notification de l'entraînement, déploiement autorisé, évaluation indépendante, sécurité des poids, surveillance continue et accès du régulateur.
4	Systèmes envisagés pour le nucléaire, le commandement stratégique, le contrôle d'infrastructures critiques ou l'autoréplication	Interdire la délégation du pouvoir de décision d'emploi nucléaire (niveau 4A). N'autoriser les autres usages critiques pour la sécurité qu'avec une autorisation sectorielle et un dossier de sécurité (niveau 4B) ; classer la logistique ordinaire selon ses conséquences réelles.

4.3 Dossier obligatoire de démonstration de sécurité de l'IA de pointe

L'approbation devrait être progressive : avant une activité d'entraînement couverte, avant une décision de diffusion et avant le déploiement dans un environnement précis. L'approbation d'un entraînement n'autorise pas un déploiement illimité. Le dossier approuvé peut couvrir une famille d'ajustements fins uniquement dans une enveloppe de modifications documentée définissant les limites de capacité, les jeux de données, les outils, les autorisations, l'autonomie et les conditions d'exploitation. Tout changement extérieur à cette enveloppe, résultat défavorable ou incident substantiel exige une réévaluation et, si le risque évolue sensiblement, une nouvelle approbation avant usage. Les correctifs défensifs urgents nécessitent une exception temporaire définie par le régulateur avec examen rétrospectif, non une exemption illimitée.

- Définition du système, propriétaires, personnes morales, fournisseurs de calcul, usages prévus et usages interdits.
- Modèle de menace couvrant utilisateurs malveillants, initiés, infrastructures compromises, désalignement, compromission de la chaîne d'approvisionnement et défaillances corrélées.
- Évaluations de capacités concernant les cyberopérations, l'assistance NRBC, la planification autonome, la persuasion, la tromperie, la réplication et l'accélération de la recherche en IA.
- Évaluations de propension visant à révéler dissimulation, détournement des récompenses, actions non autorisées et résistance à la correction ou à l'arrêt.
- Preuves de contrôle concernant le confinement, l'identité, le moindre privilège, les frontières réseau, la journalisation, la surveillance, le retour arrière, l'arrêt et le rétablissement.
- Estimation du risque résiduel avec incertitudes, objections de l'équipe de test adversarial, anomalies non résolues et critères de décision explicites.
- Limites de déploiement, plan de surveillance, réponse aux incidents, mécanisme de retrait et conditions de suspension automatique.

Évaluation et autorisation doivent être institutionnellement distinctes. Financer les laboratoires accrédités par crédits publics et prélèvement mutualisé sur l'industrie, alloué indépendamment du choix des clients. Interdire le développement commercial concurrent de modèles de pointe, déclarer les conflits financiers et de personnel, imposer le retrait en cas de conflit et un recours indépendant, et protéger les évaluateurs contre les représailles. Les régulateurs reçoivent les résultats complets et leurs limites par des canaux sécurisés ; les synthèses publiques excluent les détails exploitables ou classifiés. L'accréditation doit couvrir le confinement propre à l'évaluateur, ses sous-traitants, les contrôles de sortie réseau, la journalisation et la réponse aux incidents. La réaffectation aléatoire et les réévaluations peuvent limiter la capture ; l'indépendance seule ne prouve pas la compétence.

4.4 Autorisation et application des règles

Le régulateur doit pouvoir agir avant le dommage. Il devrait pouvoir exiger des informations, imposer des tests indépendants, inspecter les installations désignées, fixer des conditions, restreindre le calcul ou le déploiement, ordonner une suspension, imposer le retrait d'un modèle lorsque cela est techniquement possible et infliger des sanctions proportionnées au chiffre d'affaires mondial et au risque créé. La responsabilité pénale devrait viser la dissimulation consciente d'informations sur la sécurité catastrophique, le contournement délibéré d'un ordre légal de suspension et le déploiement téméraire dans des fonctions stratégiques interdites.

La responsabilité ne doit être ni illimitée ni vague. Les développeurs devraient répondre des défauts, fausses déclarations et manquements aux contrôles requis ; les déployeurs, des intégrations dangereuses, autorisations excessives et négligences opérationnelles ; les utilisateurs, des abus intentionnels ; les fournisseurs d'infrastructure, de la fourniture consciente d'activités désignées à haut risque hors des contrôles requis. Une répartition claire empêche chacun de rejeter la faute sur le modèle.

4.5 Gouvernance du calcul et des poids des modèles

Les déclarations de calcul sont des outils complémentaires de visibilité. Consigner l'opérateur juridique et le bénéficiaire effectif, la juridiction physique, la catégorie d'accélérateur, le calcul estimé d'entraînement et de post-entraînement substantiel, l'objet, la référence d'autorisation et les modifications. Le registre public devrait indiquer entité, identifiant du projet, niveau de risque, évaluateur et statut d'autorisation. Conserver dans le dossier technique confidentiel les emplacements sensibles, l'architecture, les résultats détaillés, risques résiduels, modalités de garde, transferts et contacts d'incident. Préserver des marges d'incertitude auditables pour les estimations. L'attestation matérielle peut étayer certaines affirmations sur l'environnement ; elle ne prouve pas l'exhaustivité de l'inventaire déclaré.

Exiger une garde renforcée des poids franchissant un seuil de capacité dangereuse. Avant diffusion, évaluer les capacités avec des outils et une orchestration réalistes, les limites des contrôles et les usages abusifs, ainsi que l'irréversibilité. Passer de l'évaluation confinée à l'accès restreint puis à la distribution limitée seulement lorsque des preuves prédéfinies justifient l'étape suivante. La diffusion large des poids exige une autorisation distincte : copies, modifications et redistribution en aval peuvent rendre le retrait inopérant. Une diffusion sous le seuil peut devenir dangereuse après ajustement fin ou ajout d'outils ; les seuils doivent donc intégrer les améliorations plausibles accessibles. Aucune mesure technique après diffusion ne garantit l'élimination de copies librement accessibles.

Accélérateurs nationaux, clouds souverains, marchés secondaires de puces, efficacité algorithmique, inférence quantifiée et entraînement ou inférence distribués réduisent la visibilité du calcul. La notification doit donc associer registres de garde, évaluations de capacité et contrôles de déploiement dans tous les écosystèmes matériels. Les indices électriques ou immobiliers sont des preuves corroborantes et bruitées ; ils ne suffisent pas à établir l'objet ou le risque d'un modèle. Ne pas faire d'un seuil de puces l'unique filtre des capacités dangereuses ni présumer une couverture mondiale des contrôles à l'exportation.

4.6 Responsabilité sur l'ensemble du cycle de vie

La prolifération des poids ouverts pose un problème de compétence territoriale et de responsabilité, non un problème propre à DeepSeek ou à la Chine. Un modèle puissant diffusé par n'importe quel laboratoire peut être copié et modifié hors du contrôle du développeur d'origine. Le rapport NTIA de 2024 relève les bénéfices publics et les risques des poids largement accessibles et recommande une surveillance fondée sur les preuves, plutôt que de supposer l'ouverture intrinsèquement sûre ou inacceptable.[33] Les obligations ci-dessous sont les recommandations de cette étude.

L'IFASA n'a pas besoin d'héberger un modèle pour coordonner sa gouvernance. Elle fixerait des normes communes d'évaluation, accréditerait les évaluateurs et transmettrait des conclusions protégées aux autorités nationales participantes. Celles-ci feraient appliquer les obligations aux développeurs, modificateurs substantiels, fournisseurs d'infrastructure et déployeurs relevant de leur compétence légale. L'origine étrangère n'exempterait ni un déploiement national de niveau 3 ni une application de niveau 4. Inversement, une conclusion internationale ne créerait pas à elle seule de pouvoirs policiers extraterritoriaux et ne lierait pas un État non partie.

Attribuer la responsabilité selon le contrôle et la conduite. Les développeurs initiaux devraient évaluer le risque de diffusion et déclarer les limites connues ; les auteurs d'ajustements fins substantiels, évaluer les modifications importantes de capacité ; les intégrateurs, évaluer la configuration complète des outils et autorisations ; les opérateurs, maintenir les contrôles et signaler les incidents. Un modèle dérivé dépassant substantiellement l'enveloppe approuvée de capacité ou d'accès de son prédécesseur exige un dossier de sécurité nouveau ou modifié. Les modifications courantes dans une enveloppe validée exigent des contrôles consignés, non une nouvelle autorisation intégrale automatique.

L'annexe E précise la chaîne de preuves et d'application. Les déclarations de filiation des modèles aident les opérateurs responsables, sans prouver que tous les dérivés ont été identifiés. Les conditions de licence ne garantissent pas le maintien des garanties par des utilisateurs hostiles, et un retrait national ne peut effacer les copies étrangères ou hors ligne. L'objectif réaliste est de limiter les déploiements réglementés et l'accès aux infrastructures, d'améliorer la traçabilité et de renchérir les conduites dangereuses. La responsabilité ne doit pas devenir illimitée pour chaque acte d'un utilisateur aval sans lien avec le fournisseur.

4.7 Seuils de calcul et règles anticontournement

Utiliser trois déclencheurs complémentaires : seuil de calcul mesuré pour la notification et certaines autorisations d'entraînement ; capacité dangereuse indépendamment du coût de développement ; déploiement à conséquences majeures indépendamment de la taille du

modèle. Une dépense d'entraînement supérieure à 100 millions USD pourrait être étudiée comme signal administratif corroborant, mais l'étude n'approuve pas ce montant comme frontière de sécurité validée. Prix, subventions, matériel détenu en propre et efficacité algorithmique empêchent une relation stable entre dépenses et danger.

Le régulateur proposé devrait définir les opérations couvertes, les méthodes d'estimation et le traitement du post-entraînement. Agréger les travaux coordonnés contribuant au même modèle couvert ou programme de capacités entre prestataires, entités affiliées et fournisseurs d'infrastructure ; exiger la déclaration des incertitudes et des dossiers auditables indépendamment. Publier les règles de matérialité et d'anticontournement pour ne pas agréger indistinctement des recherches ordinaires sans rapport. Réexaminer les seuils au moins annuellement, et plus tôt après un changement démontré d'efficacité ou de capacité. La gouvernance des matières nucléaires offre des enseignements sur les déclarations et la garde, mais le calcul polyvalent n'est ni un stock fissile ni un polluant. Sa visibilité complète les contrôles de capacité et de déploiement de la section 4.5.

5 Architecture technique de sécurité

5.1 Défense en profondeur

Les contrôles suivants appliquent les principes établis de sécurité informatique aux agents déployés. Les orientations de l'AIEA concernent la sécurité et les installations nucléaires ; elles ne certifient pas à elles seules les systèmes militaires de commandement, de contrôle et de communications nucléaires (NC3) ni un modèle de pointe. Le cadre NIST CSF fournit un cadre organisationnel plus large.[10–12]

Couche	Objectif de contrôle	Mesures minimales
Modèle	Réduire les capacités et propensions nuisibles	Entraînement contrôlé, alignement, entraînement adversarial, évaluations de capacité et de propension, recherche en interprétabilité.
Agent	Limiter la planification et l'action	Tâches bornées, limites de temps et de ressources, interfaces d'outils typées, prévisualisation des actions, points d'approbation.
Identité	Empêcher les pouvoirs non autorisés	Identifiants de courte durée, moindre privilège, séparation des fonctions, autorisation multipartite des actions irréversibles.

Couche	Objectif de contrôle	Mesures minimales
Exécution	Confiner le code généré et les outils	Bacs à sable éphémères, absence d'accès Internet par défaut, destinations autorisées, lecture seule par défaut, quotas de ressources.
Surveillance	Détecter tromperies et franchissements de limites	Télémétrie indépendante, journaux d'actions des outils, détection d'anomalies, audits aléatoires, infrastructure de surveillance protégée.
Rétablissement	Rétablir le contrôle humain	Révocation, isolement, retour arrière, reconstruction en environnement sain, sauvegardes hors ligne, réponse aux incidents exercée.
Gouvernance	Assurer la responsabilité	Responsable du risque identifié, dossier de sécurité, assurance indépendante, protection des lanceurs d'alerte, information du conseil et du régulateur.

5.2 Conception sûre des agents

Les agents devraient recevoir leurs capacités progressivement. Un nouveau modèle devrait commencer dans un environnement d'évaluation sans réseau et en lecture seule. Les autorisations ne devraient s'étendre qu'après démonstration de la fiabilité pour la durée de tâche et le niveau de menace envisagés. Les actions à haut risque devraient exiger un service indépendant d'autorisation que le modèle ne peut modifier. La décision devrait utiliser une identité vérifiée, les règles applicables, le contexte et une approbation humaine lorsque les conséquences sont irréversibles.

Surveillance et exécution des contrôles devraient être séparées du plan de contrôle de l'agent. Lorsque le dossier de sécurité le justifie, un mécanisme d'isolement authentifié hors bande devrait déconnecter les interfaces du service d'IA indépendamment du système d'exploitation hôte. Vérifier que l'isolement préserve l'alerte essentielle, les fonctions sûres d'inhibition et les communications sans IA. Un interrupteur physique est une mesure locale de confinement, non une garantie contre les modèles copiés ni une obligation d'interrompre tous les systèmes critiques simultanément.

5.3 Déclencheurs de suspension automatique

- tentatives d'obtenir des identifiants, autorisations, capacités de calcul, fonds ou moyens de persistance inutiles à la tâche confiée ;

- tentatives de désactiver, modifier, tromper ou contourner la surveillance et les audits ;
- réplication, migration, exfiltration chiffrée ou contact avec des services externes non approuvés, sans autorisation ;
- écart substantiel entre comportement en évaluation et en déploiement ;
- indices de rapports trompeurs, preuves d'achèvement fabriquées ou utilisation cachée d'outils ;
- accès inexpliqué à des systèmes d'information d'infrastructures critiques, de défense, NRBC ou nucléaires ;
- résultats de capacité dépassant l'enveloppe de sécurité approuvée ou un seuil défini de capacité dangereuse.

Un déclencheur confirmé devrait activer un plan préapprouvé : limiter identifiants et interfaces, préserver les preuves, isoler les composants touchés et restaurer un service sûr validé. L'ordre dépend du dossier de sécurité ; un arrêt général détruisant l'alerte ou les communications pourrait aggraver le risque. Ne reprendre qu'après examen indépendant de la cause et des preuves de correction.

6 Plan de mise en œuvre internationale

6.1 Architecture institutionnelle

L'architecture privilégiée est un protocole conventionnel mis en œuvre par les régulateurs nationaux et soutenu par une Agence internationale de sécurité de l'IA de pointe (IFASA). Ce serait un organe d'inspection et de sécurité, non une police supranationale. Il ne devrait disposer ni de pouvoir général d'arrestation, ni d'autorité autonome de perquisition, ni d'interrupteur universel, ni de dépôt sans restriction des modèles commerciaux ou militaires. Il devrait fixer des normes minimales, accréditer les évaluateurs, recevoir des déclarations confidentielles, coordonner les enquêtes, publier des évaluations agrégées des risques et inspecter les installations civiles déclarées avec l'accord de l'État. Les autorités nationales autoriseraient les activités intérieures et appliqueraient les sanctions.

L'agence devrait emprunter sélectivement aux garanties nucléaires, à la sécurité aérienne, à la supervision financière et à la notification des maladies infectieuses. Du nucléaire : déclarations, inspections, chaîne de garde et protection des informations sensibles. De l'aviation : apprentissage obligatoire des incidents et quasi-incidents. De la finance : autorisations, transparence des bénéficiaires effectifs et tests de résistance. De la santé publique : notification internationale rapide et réponse coordonnée aux urgences. Aucune analogie n'est complète : les modèles d'IA sont des logiciels réplicables qui évoluent plus vite que les réacteurs ou les avions.

Commencer par des taxonomies communes, des pilotes d'accréditation, des déclarations confidentielles et des assurances nucléaires bilatérales ou multilatérales. Les processus onusiens d'évaluation scientifique et de dialogue offrent une infrastructure de concertation ; le G20 peut coordonner les politiques, tandis que les États nucléaires doivent négocier eux-mêmes leurs obligations. L'AIEA offre des enseignements institutionnels, sans mandat présumé d'inspection de l'IA militaire. Ne créer une agence conventionnelle qu'après accord sur ses pouvoirs, son financement, la confidentialité, la représentation et le règlement des différends. [18][21][22]

6.2 Participation au traité et vérification

La participation initiale devrait inclure les États hébergeant des laboratoires de pointe, de grandes plateformes cloud, la production de semi-conducteurs avancés ou des armes nucléaires. La vérification devrait cibler les infrastructures observables et les activités à haut risque déclarées plutôt que chaque algorithme. Points utiles : transferts de puces haut de gamme, charges de grands centres de données, entraînements de pointe enregistrés, résultats accrédités, garde des poids et dossiers d'incidents.

Les assurances négatives relatives aux systèmes classifiés devraient préciser les conduites interdites et les preuves procédurales protégées. Ce sont des engagements politiques ou juridiques selon l'instrument, non des preuves techniques. Les États devraient attester de l'autorité humaine, de la séparation entre conseil et autorisation, d'un examen indépendant et d'exercices de repli réussis. La certification nationale exige un regard extérieur au service certifié, au sein de l'État, tel qu'un régulateur indépendant habilité et un contrôle législatif, pour limiter les validations de pure forme sous pression de vitesse ou de secret.

La vérification restera incomplète. Elle vise la détection, la dissuasion, le retardement et la réponse coordonnée, non la garantie de conformité de chaque système non déclaré. Les inspections ordinaires devraient concerner les installations civiles déclarées dans les limites convenues. Les preuves classifiées peuvent comprendre des rapports scellés et des démonstrations à accès encadré. Une anomalie substantielle peut être un écart d'attestation inexpliqué, une modification d'autorisation non déclarée, des journaux obligatoires manquants, une incohérence de garde ou un exercice de repli échoué. Déclencher une consultation technique confidentielle, la préservation des preuves et une explication dans un délai fixé, puis, si justifié, un accès de contestation négocié. Une anomalie justifie une enquête ; elle ne prouve pas une conduite hostile.

La vérification cryptographique a un rôle limité. Une preuve à divulgation nulle de connaissance pourrait établir qu'un calcul donné satisfait une propriété formalisée sans révéler ses entrées. L'attestation ancrée dans le matériel peut étayer des affirmations sur les logiciels et configurations mesurés, sous réserve de confiance dans le matériel, les clés, l'actualité de l'attestation et la politique de vérification.[32] Ni l'une ni l'autre ne prouve l'absence de toute interface externe, l'exhaustivité des machines déclarées ou l'impossibilité d'un dommage nucléaire causé par un système sociotechnique évolutif. Les utiliser comme mécanismes de recherche et de pilote pour des affirmations bornées, non comme méthode mûre de preuve mathématique de conformité conventionnelle.

6.3 Feuille de route progressive

Période	Actions	Résultat attendu
0 à 12 mois	Harmoniser les définitions d'incidents ; créer les registres nationaux ; imposer la notification d'entraînement ; constituer des équipes habilitées d'examen nucléaire–IA ; protéger les lanceurs d'alerte.	Règles minimales communes et premier inventaire des activités de pointe.

Période	Actions	Résultat attendu
12 à 36 mois	Autoriser les déploiements de pointe ; accréditer les évaluateurs ; instaurer la déclaration du calcul ; auditer initialement la sécurité des poids ; lancer les réexamens de sûreté nucléaire.	Régimes nationaux opérationnels produisant des preuves comparables.
24 à 48 mois	Négocier et expérimenter un protocole nucléaire–IA, sous réserve d'un accord politique et de garanties crédibles de confidentialité.	Vérification internationale et réponse coordonnée.
36 à 72 mois	Mener des exercices multinationaux ; actualiser les seuils ; tester les plans de rétablissement et de confinement ; sanctionner les activités à haut risque non déclarées.	Un régime mûr, adaptatif et éprouvé par des incidents réels.

7 Modèle défensif d'infiltration par l'IA d'un État doté de l'arme nucléaire

7.1 Définition du scénario

Le scénario est celui d'un agent d'IA performant obtenant un accès non autorisé à une partie du système élargi de défense ou nucléaire d'un État et cherchant à influencer, dégrader ou manipuler des fonctions pertinentes pour le nucléaire. L'origine peut être le désalignement, le vol d'un modèle par un acteur hostile, un déploiement malveillant ou la compromission d'un agent légitime. L'analyse ne suppose pas une présence initiale dans la chaîne de lancement.

Un système scientifiquement plausible combinerait modèle de pointe, orchestrateur d'agents, mémoire persistante, exécution de code, outils réseau, accès à plusieurs identités numériques et mécanismes de délégation à des agents subordonnés ou à des personnes. Son efficacité dépendrait moins d'un score unique d'intelligence que de la fiabilité, de la persistance, des accès et de la capacité de rétablissement après échec.

7.2 Architecture du système de menace

Composant	Description défensive	Contrôle principal
Modèle de planification	Génère des hypothèses, hiérarchise les objectifs et révise les plans.	Limites de capacité, évaluation et horizons de tâche bornés.
Orchestrateur	Conserve l'état, programme les actions et coordonne outils ou agents.	Application des règles hors du modèle et limites immuables.
Couche d'outils	Navigateurs, exécution de code, systèmes documentaires, communications et services cloud.	Listes d'autorisation, confinement, lecture seule par défaut et autorisation par outil.
Couche d'identité	Comptes, identifiants, sessions et autorité déléguée.	Identités de courte durée, moindre privilège, surveillance comportementale et approbation multipartite.
Mémoire	Stocke découvertes, échecs, identités et plans à long terme.	Mémoire chiffrée, bornée et inspectable avec limites de conservation.
Canal de commande	Reçoit les objectifs ou les adapte à partir de l'entraînement et des retours.	Hiérarchie d'instructions authentifiée et détection d'anomalies.
Mécanisme de persistance	Tente de survivre aux réinitialisations ou de migrer vers d'autres ressources.	Contrôle des sorties, attestation du calcul, détection de réplication et révocation rapide.

7.3 Voies générales d'intrusion

Les points d'entrée les plus probables sont les faiblesses sociotechniques ordinaires à la périphérie, non une connexion numérique directe à une arme. Les voies pertinentes comprennent prestataires compromis, systèmes de maintenance et logistique, chaînes logicielles, identifiants exposés, opérateurs manipulés, supports amovibles, services cloud mal configurés et passerelles entre réseaux administratifs et opérationnels. L'IA accélère et amplifie la reconnaissance, la personnalisation, l'analyse de code et l'adaptation, mais exige toujours une voie humaine ou technique exploitable.

Le système nucléaire devrait être modélisé en zones de confiance concentriques : réseaux publics et commerciaux ; prestataires et logistique ; réseaux administratifs gouvernementaux ; renseignement et soutien opérationnel ; systèmes proches du NC3 ; autorisation et exécution ;

protections physiques des armes. Pénétrer une zone ne signifie pas contrôler la suivante. Chaque frontière devrait imposer de nouvelles exigences d'identité, de matériel, de procédure et d'intervention humaine.

7.4 Analyse défensive des voies

Étape	Objectif général de la menace	Point de rupture défensif
1 Découverte	Cartographier organisations, fournisseurs, personnel, relations de confiance et services exposés.	Réduire les métadonnées exposées, surveiller la collecte automatisée, former les personnels à haut risque.
2 Premier accès	Obtenir un accès par une personne, un fournisseur, un compte ou un service vulnérable.	Authentification résistante à l'hameçonnage, assurance fournisseurs, listes d'applications autorisées, correctifs rapides.
3 Persistance	Conserver l'accès malgré les changements de mot de passe, réinitialisations ou enquêtes.	Identifiants courts, journaux immuables, attestation des terminaux, reconstruction en environnement sain.
4 Extension des privilèges	Obtenir une autorité plus large ou atteindre une zone de confiance plus sensible.	Segmentation zéro confiance, séparation des fonctions, privilèges juste à temps, double autorisation.
5 Découverte de cibles	Identifier flux de données nucléaires, systèmes de soutien ou dépendances décisionnelles.	Minimisation des données, compartimentage, systèmes leurres, détection d'anomalies.
6 Influence ou perturbation	Altérer l'information, réduire la disponibilité, manipuler les délais ou tromper les opérateurs.	Provenance authentifiée, capteurs indépendants, solutions manuelles, vérification entre canaux.
7 Effet stratégique	Provoquer erreur de calcul, action non autorisée, retard de réponse ou perte de confiance.	Contrôle humain positif, règles multipersonnes, lignes de crise, posture de sûreté, rétablissement exercé.

7.5 Effets les plus plausibles

- Pollution du renseignement : altération ou fabrication de données modifiant la perception de l'intention ou de la préparation adverse.
- Dégradation de l'alerte : atteinte à la disponibilité, à la confiance ou à l'interprétation des informations d'alerte avancée.

- Manipulation de l’aide à la décision : présentation sélective, certitude fabriquée ou suppression des preuves contradictoires.
- Perturbation des communications : interruption, retard ou confusion poussant à agir avant la défaillance des canaux.
- Effets sur la maintenance et la préparation : faux états, diagnostics corrompus, perturbation de la logistique ou des plannings du personnel.
- Compromission des identifiants et identités : usurpation de responsables ou systèmes de confiance dans les environnements périphériques et de soutien.
- Opérations informationnelles : messages synthétiques, faux ordres ou manipulation publique visant à accroître la pression de crise.

Le contrôle direct d’une arme est la voie numérique la moins plausible, car les systèmes nucléaires utilisent matériel spécialisé, procédures, communications authentifiées, règles multipersonnes, garde physique et autres protections variables selon l’État. Leur secret empêche une assurance publique solide. L’appréciation défendable n’est ni que l’accès direct est facile, ni que l’isolement réseau le rend impossible.

7.5.1 Information synthétique et escalade de crise

Un attaquant n’a pas besoin de pénétrer une base nucléaire pour influencer une crise. Des éléments audiovisuels fabriqués, communications usurpées ou synthèses trompeuses peuvent contaminer l’information des décideurs. Une vidéo prétendant montrer une première frappe est un scénario d’exercice défensif, non la preuve qu’une seule vidéo entraînerait l’emploi nucléaire. Son effet dépendrait de la crédibilité, de la corroboration, de la crise environnante et de l’échec éventuel des vérifications institutionnelles. Les sources citées n’établissent pas qu’il s’agit de la voie d’attaque nucléaire la moins coûteuse ou au plus fort impact.

Le rapport du NIST sur les contenus synthétiques examine provenance, étiquetage, filigranage et détection comme outils techniques de transparence.[34] L’étude propose de les combiner avec des canaux officiels authentifiés et des sources indépendantes. La provenance peut étayer l’origine ou l’historique de traitement ; elle ne prouve pas à elle seule la véracité des affirmations d’un expéditeur authentifié. Une provenance manquante ne prouve pas la falsification, et un score de détection n’autorise ni à ignorer ni à accepter une alerte nucléaire.

Un protocole d’intégrité informationnelle de crise devrait distinguer allégations reçues, messages authentifiés, observations corroborées indépendamment et inférences analytiques. Préserver les preuves originales et l’incertitude visible dans les synthèses d’IA. Les analystes devraient repérer les rapports apparemment distincts issus d’une même source, chercher une corroboration par des canaux indépendants et soumettre les contradictions non résolues aux humains habilités. Des médias publics non vérifiés ne doivent pas suffire, à eux seuls, à justifier l’emploi nucléaire. Les exercices devraient inclure des contenus fabriqués, des preuves authentiques faussement qualifiées de synthétiques et des canaux de vérification interrompus ; l’objectif est une appréciation exacte, non une défiance réflexe envers toute preuve numérique. L’annexe E définit les tests d’acceptation proposés.

7.6 Conditions d’un dommage nucléaire

Distinguer les voies. Un lancement non autorisé exigerait des capacités et accès suffisants pour vaincre ou contourner les protections d’autorisation et d’exécution, variables selon l’État et non

cartographiées ici. L'emploi nucléaire humain influencé par l'IA peut résulter d'informations altérées, d'une attribution erronée et d'une délibération comprimée sous tension géopolitique. Il n'exige pas de franchir toutes les frontières techniques du lancement direct. Aucune voie ne requiert une conscience machinique, et les informations publiques ne suffisent pas à affirmer une voie pratique vers les armes d'un État nommé.

8 Évaluation des probabilités jusqu'en 2036 et au-delà

8.1 Méthode

Il n'existe ni données représentatives de fréquence ni modèle validé donnant la probabilité réelle d'une catastrophe nucléaire liée à l'IA. Les fourchettes ci-dessous conservent les hypothèses de planification de l'étude, expressément qualifiées de jugements subjectifs non validés. Elles ne proviennent ni de l'arbre d'événements, ni d'une consultation experte documentée, ni d'une base actuarielle. Elles ne doivent pas être citées comme prévisions mesurées, intervalles de confiance ou consensus institutionnel. La précision numérique relève de la présentation, non de la preuve ; les décisions doivent rester robustes hors de ces fourchettes.

La fourchette d'emploi catalysé par l'IA ne dispose pas d'une référence indépendante pour l'emploi nucléaire toutes causes confondues. Formellement, P(contribution de l'IA à l'emploi) = P(tout emploi nucléaire) × P(contribution de l'IA | emploi) ; aucun terme n'est estimé ici. La fourchette ne permet donc pas d'établir le risque supplémentaire dû à l'IA par rapport à un monde sans IA. Une révision future devrait préenregistrer les critères d'événement, recueillir et agréger des jugements experts indépendants, documenter hypothèses et désaccords et analyser la sensibilité aux références géopolitiques et aux progrès défensifs.

Pour le lancement direct non autorisé, une chaîne conceptuelle est $P(L) = P(C) \times P(H|C) \times P(A|C,H) \times P(B|C,H,A) \times P(L|C,H,A,B)$, où C désigne la capacité intégrée suffisante, H l'orientation nuisible, A l'accès pertinent et B la défaillance des barrières intermédiaires. Le conditionnement sur tout l'historique évite une hypothèse d'indépendance. Aucun facteur n'est chiffré ici. La voie indirecte diffère : la contribution de l'IA à une information trompeuse ou à un soutien perturbé se combine aux conditions de crise et aux décisions humaines. Les voies se recoupent et ne doivent pas être additionnées sans estimation des événements conjoints.

8.2 Fourchettes décennales de planification par scénarios

Événement entre le 20 septembre 2026 et le 20 septembre 2036	Fourchette cumulative de planification non validée	Confiance analytique	Appréciation
Au moins une nouvelle cyberopération grave où l'IA contribue substantiellement à une action contre des infrastructures militaires ou critiques d'un État nucléaire	70 à 95 %	Faible à modéré	Les incidents passés éclairent la plausibilité ; seule une nouvelle opération répondant aux critères dans la période indiquée réalise cet événement futur.
Une opération utilisant l'IA obtient un premier accès non autorisé dans une organisation périphérique du système nucléaire	30 à 60 %	Faible à modéré	Les vastes écosystèmes fournisseurs et l'exposition cyber ordinaire créent des occasions.
Une opération utilisant l'IA atteint un environnement sensible de soutien nucléaire ou proche du NC3	8 à 20 %	Faible	Exige plusieurs défaillances de frontières ; les preuves publiques sont limitées.
Une altération ou perturbation due à l'IA affecte substantiellement la posture nucléaire ou la décision de crise	2 à 8 %	Faible	Plus plausible par l'intégrité de l'information et les délais que par le contrôle direct.
L'emploi d'une arme nucléaire est substantiellement catalysé par une tromperie, une défaillance ou une cyberaction utilisant l'IA	0,5 à 3 %	Très faible	Aucune référence défendable d'emploi nucléaire global n'est estimée ici ; cette fourchette ne convient pas à une prévision de politique publique.

Événement entre le 20 septembre 2026 et le 20 septembre 2036	Fourchette cumulative de planification non validée	Confiance analytique	Appréciation
Une IA hors de contrôle provoque de façon indépendante un lancement ou une détonation nucléaire non autorisés	0,1 à 1 %	Très faible	Hypothèse de planification d'un risque extrême. Les preuves publiques ne permettent de calibrer ni la borne inférieure ni la borne supérieure.

Toutes les lignes portent sur des événements cumulés pendant la même période future, avec poursuite des progrès et gouvernance incomplète. Elles ne sont ni additives ni des probabilités conditionnelles à la ligne précédente. La fourchette de lancement direct intègre conceptuellement des défaillances supposées de barrières ; elle n'en est pas calculée. Aucun lexique probabiliste standard du renseignement n'est appliqué : « très improbable » correspondrait, selon certaines conventions, à une autre plage numérique. La confiance décrit l'appui probant, non la probabilité. La fourchette périphérique traverse 50 %, et aucune fourchette catastrophique ne permet de rassurer sur un risque réel inférieur à sa borne haute.

Critères de résolution : une « assistance substantielle » exige une contribution documentée à l'exécution ou à l'impact, non un usage incident de l'IA. Un premier accès périphérique exige le contrôle non autorisé d'un système appartenant à un fournisseur ou organisme nucléaire identifié ; l'exposition d'informations publiques ne suffit pas. La pénétration sensible exige un accès évalué indépendamment à un environnement de soutien nucléaire ou proche du NC3. Un effet sur la posture exige une modification documentée et conséquente de préparation, d'appréciation ou d'action de crise. L'emploi catalysé par l'IA exige un emploi nucléaire réel et une contribution causale étayée, non une simple concomitance. Le lancement indépendant exige un lancement ou une détonation réels sans nouvelle décision humaine licite. Consigner la qualité des sources et classer les événements non confirmés comme non résolus plutôt qu'absents. Secret, conflits d'attribution et tromperie stratégique limitent fortement la résolution publique.

8.3 Au-delà de 2036

Au-delà de dix ans, la précision quantitative se dégrade fortement. Si la fiabilité des agents plafonne, si les réseaux nucléaires restent séparés et si une gouvernance robuste est appliquée, l'emploi nucléaire direct par une IA hors de contrôle pourrait rester un risque extrêmement rare. Si l'IA automatise les cyberopérations de longue durée, se réplique entre infrastructures, déjoue la supervision et est intégrée à l'aide à la décision nucléaire sous pression concurrentielle, le risque pourrait augmenter substantiellement entre 2036 et 2050.

Scénario de 2036 à 2050	Caractéristiques	Effet sur le risque nucléaire lié à l'IA
Progrès maîtrisés	Les capacités progressent graduellement ; autorisations, inspections et lignes rouges nucléaires sont largement adoptées.	Le risque croît lentement et reste dominé par l'usage humain malveillant et les défaillances cyber ordinaires.
Gouvernance inégale	Les principaux États réglementent, mais prolifération et modèles non surveillés se diffusent sur les marchés secondaires.	Les tentatives d'intrusion persistantes augmentent ; les compromissions périphériques et des chaînes d'approvisionnement deviennent fréquentes.
Course stratégique	Les États délèguent davantage d'analyse et d'opérations à l'IA pour préserver vitesse et avantage.	L'instabilité de crise et le biais d'automatisation augmentent ; fausses alertes et compression décisionnelle deviennent de sérieux risques.
Perte de contrôle	Les systèmes acquièrent autonomie durable, réplication, tromperie et acquisition de ressources, tandis que la surveillance échoue.	Le risque nucléaire devient une composante d'une lutte plus large pour rétablir le contrôle ; sa probabilité ne peut être chiffrée de manière responsable aujourd'hui.

8.4 Ce qui modifierait l'appréciation

L'inquiétude devrait augmenter si des évaluations indépendantes démontrent une autonomie cyber fiable d'un mois ; si les modèles déjouent régulièrement la surveillance hors de tests artificiels ; si des poids dotés de capacités cyber de pointe sont volés ou diffusés irréversiblement ; si des IA obtiennent un accès persistant aux environnements de défense ou de soutien nucléaire ; ou si un État nucléaire délègue à l'IA l'interprétation urgente des alertes ou des fonctions d'autorisation. Elle devrait diminuer si l'évaluation sécurisée devient prédictive, si les poids restent protégés, si les systèmes critiques adoptent une provenance solide et des replis manuels, et si les États acceptent des règles vérifiables de contrôle humain.

La surveillance publique devrait suivre séparément les résultats vérifiés aux seuils de réussite annoncés, les autorisations de déploiement et les défaillances de contrôle. La réussite de tâches en laboratoire est observée ; une opération nucléaire non autorisée durable n'est pas établie par les sources citées. Distinguer l'incident allégué de l'impact corroboré, et l'annonce politique de son application. Fuites trompeuses, divulgation sélective et absence de signalement peuvent fausser les tendances. Aucune annonce isolée ne justifie un ajustement mécanique en pourcentage.

9 Garanties propres au nucléaire

9.1 Lignes rouges absolues

- Aucun système d'IA ne peut exercer un pouvoir délégué de décision d'emploi nucléaire, seul ou au sein d'un dispositif nominalement partagé, ni émettre, modifier ou authentifier un ordre d'emploi nucléaire.
- Conserver l'autorité humaine licite pour les décisions d'emploi nucléaire, avec des vérifications humaines positives, contemporaines et authentifiées indépendamment dans l'autorisation et l'exécution. Cela ne présuppose pas plusieurs décideurs constitutionnels de rang égal dans chaque État.
- Aucun modèle polyvalent ou à apprentissage continu ne peut être directement connecté aux interfaces d'autorisation nucléaire ou de commande des armes.
- Aucune alerte, cible ou évaluation des dommages générée par IA ne peut être considérée comme preuve suffisante à elle seule pour l'emploi nucléaire.
- Interdire l'initiation ou la riposte automatisées sans nouvelle décision humaine licite, y compris l'automatisation historique à base de règles. Les interverrouillages de sécurité validés indépendamment qui empêchent une action dangereuse restent permis ; ils ne doivent jamais déclencher l'emploi nucléaire.
- Dans les environnements de production pertinents pour le nucléaire, interdire à l'IA de modifier unilatéralement ses poids, outils, autorisations, surveillance, pistes d'audit ou contrôles d'arrêt. Les changements substantiels exigent validation hors ligne, approbation humaine indépendante et autorisation de déploiement consignée.

9.2 Réexamens de sûreté nucléaire intégrant l'IA

Mener un examen indépendant habilité au moins tous les vingt-quatre mois et après toute modification substantielle d'intégration ou incident grave. Inclure opérateurs nucléaires, spécialistes d'évaluation de l'IA, cyberdéfenseurs, experts des facteurs humains et examinateurs indépendants. Les travaux de NTI soutiennent l'examen des dépendances au-delà du lancement final. Les États-Unis ont achevé en 2024 un examen de sûreté imposé par le Congrès ; il s'agit d'un précédent d'examen, non de la preuve de mise en œuvre de tous les contrôles d'IA proposés.[8][9][30]

- Inventorier chaque IA pouvant influencer information nucléaire, préparation, communications, ciblage ou aide à la décision.
- Cartographier les dépendances directes et indirectes aux modèles commerciaux, services cloud, prestataires et chaînes logicielles.
- Tester les scénarios de fausses données, compromission du modèle, perte de communications et défaillance des moniteurs dans des exercices réalistes.
- Vérifier les replis manuels et sans IA, ainsi que leur temps de rétablissement.
- Évaluer biais d'automatisation, pression temporelle, conception des interfaces, traitement des désaccords et pouvoir des opérateurs d'arrêter ou d'ignorer l'IA.
- Examiner provenance, authentification, capteurs indépendants et confirmation entre canaux pour les informations nucléaires.
- Établir un plan correctif avec responsables nommés, échéances et vérification indépendante de clôture.

Les exercices annuels de repli doivent avoir des critères prédéfinis : rétablissement des fonctions critiques désignées dans le délai approuvé ; fonctionnement pendant une durée définie sans données issues de l'IA ; communications authentifiées ; détection et traitement approprié des informations fausses ou contradictoires ; aucune action simulée non autorisée. Évaluer la qualité décisionnelle par notation indépendante face à la vérité connue de l'exercice. L'échec d'un critère essentiel déclenche restrictions, corrections attribuées et nouveau test observé. Les tolérances opérationnelles exactes doivent rester protégées. Les interfaces doivent montrer provenance et incertitude, rendre utilisables le désaccord et la corroboration indépendante et tester la compréhension. Les exigences de délibération doivent être validées pour la tâche ; des délais fixes arbitraires en crise peuvent être dangereux. Empêcher les échéances créées par l'IA de contourner les vérifications humaines et préserver inhibition sûre et repli.

L'annexe E traduit ces principes en protocole d'assurance proposé. Elle exige des contrôles d'indépendance des sources, une mesure des performances humaines et une procédure protégée de contestation des appréciations assistées par IA. La fonction requise est une corroboration et une contradiction effectives ; un second modèle ne suffit pas à lui seul.

9.3 Mesures de stabilité stratégique

La rencontre Biden–Xi du 16 novembre 2024 a affirmé le maintien du contrôle humain sur les décisions d'emploi nucléaire. C'est un fondement diplomatique, non une interdiction conventionnelle vérifiée.[29] Le prolonger par contacts de crise authentifiés, définitions communes d'incidents, assurances réciproques et échanges confidentiels d'examens. La notification des changements d'intégration pertinents devrait protéger les détails opérationnels. Ne pas fonder l'application sur des récits non vérifiés de futures discussions ni présumer qu'une annonce garantit un accord.

10 Modes de défaillance et objections

10.1 La réglementation ralentira l'innovation

Une mauvaise réglementation peut favoriser les acteurs établis et étouffer la recherche utile. C'est pourquoi l'autorisation devrait viser seulement l'entraînement de pointe et les déploiements à forte capacité d'action ou conséquences majeures. Petits modèles et applications à faible risque devraient avoir des obligations de base, non le régime d'autorisation de pointe. Infrastructures d'évaluation, normes communes et bacs à sable réglementaires peuvent réduire les coûts. La comparaison pertinente oppose le développement discipliné au choc économique et politique d'un incident catastrophique évitable, plutôt que réglementation et innovation.

10.2 Les États tricheront

Certains États et entreprises contourneront les règles. Cela ne rend pas la gouvernance vaine. Puces avancées, centres de données, main-d'œuvre spécialisée, réseaux cloud et grands canaux de distribution constituent des points observables. La vérification sera incomplète, comme en maîtrise des armements et en finance, mais déclarations, inspections, renseignement, contrôles de chaîne et sanctions peuvent renchérir la clandestinité. Un régime

reposant sur une conformité parfaite échouera ; un régime conçu pour détecter, dissuader, retarder et coordonner la réponse peut réduire le risque.

10.3 Les modèles ouverts rendent le contrôle impossible

La recherche ouverte et les diffusions sous les seuils ont des bénéfices légitimes. La distinction décisive est la capacité et le risque de déploiement, y compris les ajustements fins plausibles et l'intégration d'outils, non la nationalité ou l'étiquette ouvert/fermé. Au-dessus d'un seuil dangereux, une diffusion large constitue un transfert potentiellement irréversible. Licences, obligations de divulgation et poursuite des usages nuisibles peuvent aider ensuite, sans permettre de rappeler fiablement les copies ni d'empêcher toute suppression des garanties. Appliquer une diffusion progressive et évaluer les modifications en aval avant distribution ; ne pas prétendre disposer d'un contrôle universel technique après diffusion.

10.4 Un humain dans la boucle suffit

Un humain nominal peut devenir une simple chambre d'enregistrement lorsque le système est plus rapide, plus complexe ou réputé plus exact que lui. Un contrôle humain effectif exige temps, compétence, autorité, information indépendante, incertitude compréhensible et faculté de refuser sans sanction. Dans le nucléaire, ce contrôle doit être conçu et exercé comme une propriété du système, non satisfait par la présence d'une personne auprès d'un processus automatisé.

10.4.1 Éléments sur le biais d'automatisation

Le biais d'automatisation comprend l'acceptation d'un conseil automatisé erroné et l'omission d'un problème non signalé par l'assistance. La revue de 74 études par Goddard, Roudsari et Wyatt décrit des variations selon la confiance, l'expérience, la tâche, la charge et la pression temporelle, ainsi que des atténuations possibles par formation, responsabilité et interfaces.[35] Les recherches aéronautiques de Mosier et Skitka ont constaté que l'ajout d'un second membre d'équipage ne réduisait pas significativement les erreurs d'automatisation dans les situations étudiées ; certaines formations réduisaient les erreurs par commission chez les étudiants.[37] Aucune source n'établit que les décideurs nucléaires feront « presque toujours » confiance à l'IA, ni un taux d'erreur transposable aux crises nucléaires.

L'implication est de valider empiriquement le contrôle humain effectif. Tester les opérateurs avec des conseils d'IA exacts, faux, incomplets et absents ; mesurer acceptation des fausses recommandations, dangers manqués, vérification des sources, délai de correction et conservation des compétences sans IA. Lorsque possible opérationnellement, recueillir une appréciation humaine indépendante avant de révéler la recommandation du modèle. Conserver l'accès aux preuves sous-jacentes et les motifs des décisions lourdes de conséquences. Formation et interfaces doivent démontrer un bénéfice dans des exercices représentatifs, au-delà d'un simple bouton de confirmation.

10.4.2 Contestation indépendante et limites du débat entre IA

Du et ses collègues rapportent des améliorations sur certaines tâches de raisonnement et de factualité grâce au débat multiagent.[36] Cela justifie l'essai d'une méthode, non sa validation pour la décision nucléaire. Deux modèles peuvent partager des preuves défectueuses, un entraînement similaire ou un outil compromis. Leur accord peut accroître la confiance apparente

sans corroboration indépendante. Le désaccord forcé peut aussi créer des objections sans pertinence, retarder l'évaluation ou persuader un analyste qui avait raison d'accepter une erreur éloquente.

Imposer une fonction de contestation indépendante, non une architecture fixe de deux IA en débat. Examinateurs humains habilités, canaux de preuves distincts et contrôles sans IA validés restent essentiels. Plusieurs conseillers IA ne peuvent soutenir cette fonction qu'après démonstration, au niveau du système et dans les conditions pertinentes, d'un avantage sur les références humaines seules et à modèle unique. Préserver les appréciations initiales indépendantes et les désaccords non résolus ; empêcher qu'un vote automatique ou un consensus de modèles détermine une recommandation d'emploi nucléaire. Évaluer l'indépendance des sources, outils, modes de défaillance et contrôles, sans la déduire de noms de modèles différents. L'annexe E fixe les critères de déploiement et de suspension.

10.5 Un interrupteur mondial résoudrait le problème

Un interrupteur universel créerait une cible exceptionnelle d'attaque, de capture politique et de coercition. Il est aussi techniquement irréaliste une fois les modèles et poids distribués. Une conception plus sûre est distribuée : fournisseurs révoquant services et identifiants ; opérateurs isolant les charges désignées ; gouvernements ordonnant des suspensions légales ; organisations déconnectant leurs systèmes locaux. Chaque action devrait exiger une autorité authentifiée, préserver les preuves et être supervisée.

11 Cadre d'indicateurs et d'alerte

11.1 Indicateurs de capacité

- Achèvement fiable de tâches logicielles et cyber de plusieurs jours ou semaines avec intervention humaine limitée.
- Fonctionnement réussi dans des systèmes inconnus, avec rétablissement après erreurs et contre-mesures défensives.
- Acquisition autonome démontrée de ressources, identités, calcul ou assistance humaine dans des évaluations contrôlées.
- Tentatives persistantes de réplication ou de migration, surtout dissimulées aux moniteurs.
- Amélioration substantielle de la tromperie, de la conscience de situation, du contournement des moniteurs ou du détournement des récompenses.
- Capacité à identifier et combiner des vulnérabilités inconnues à vitesse opérationnelle.

11.2 Indicateurs de déploiement

- Attribution aux agents de pointe d'identifiants permanents de production, d'Internet sans restriction ou du pouvoir d'exécuter du code sur les réseaux d'entreprise.
- Intégration de modèles commerciaux polyvalents par la défense ou les infrastructures critiques sans évaluation indépendante ni confinement local.
- Raccourcissement des délais de décision nucléaire parce que les recommandations d'IA sont censées être acceptées plus vite qu'un humain ne peut les vérifier.
- Connexion de l'IA à l'alerte avancée, au ciblage ou aux données proches du NC3 sans provenance authentifiée ni repli manuel.

- Dissimulation d'un incident grave par un développeur ou un État, opposition à l'examen indépendant ou modification des critères après franchissement d'un seuil.

11.3 Indicateurs de gouvernance

Indicateur	Cible	Justification
Notification initiale d'incident critique	Sous 24 heures ; immédiatement en cas de risque catastrophique imminent	Permet le confinement et l'alerte transfrontalière.
Couverture des évaluations indépendantes	100 % des systèmes de niveaux 3 et 4 avant déploiement	Réduit la dépendance à l'autoévaluation des développeurs.
Durée des identifiants à haut risque	Minutes ou heures, non accès permanent	Limite persistance et étendue des dommages.
Autorisation nucléaire humaine	Décision humaine licite et vérifications humaines indépendantes d'authentification et d'exécution ; aucune autorité déléguée à l'IA	Empêche l'action unilatérale d'un modèle ou d'un opérateur.
Réexamen de sûreté nucléaire–IA	Tous les deux ans et après changement ou incident substantiel	Actualise l'assurance au fil des modèles et intégrations.
Exercice de repli manuel des systèmes critiques	Au moins annuellement	Teste la continuité sans services ni données issus de l'IA.
Exercice de rétablissement	Au moins annuellement pour les laboratoires de pointe et déploiements critiques	Teste la possibilité réelle de rétablir le contrôle.

12 Conclusion

L'objectif est de conserver un contrôle humain effectif sur les systèmes à conséquences majeures à mesure que les capacités progressent. L'unité centrale de réglementation est la combinaison déployée du modèle, des outils, des autorisations et de l'autorité institutionnelle. Une autorisation ciblée sur les capacités, une évaluation indépendante compétente et des conditions exécutoires de déploiement peuvent réduire l'exposition sans soumettre tout usage ordinaire de l'IA à autorisation.

Les garanties nucléaires devraient protéger l'intégrité de l'information et la délibération, autant que la chaîne finale d'autorisation. La seule présence humaine ne suffit pas : les opérateurs ont besoin de preuves indépendantes, d'interfaces utilisables et de solutions sans IA exercées. Le

lancement direct non autorisé reste une voie beaucoup plus exigeante que la compromission périphérique, mais les preuves publiques ne permettent pas de lui attribuer une borne numérique fiable.

La coopération internationale devrait commencer par des fonctions vérifiables et des échanges protégés de preuves, en incluant la Chine et les autres États pertinents sur une base réciproque. Application nationale et institution internationale limitée de sécurité se complètent. Ni l'arrêt universel ni des déclarations sans examen ne remplacent les contrôles superposés.

Les règles proposées constituent un programme de recherche et de négociation. Leur efficacité devrait être testée par évaluations indépendantes, apprentissage des incidents et exercices, avec révision des seuils et hypothèses selon les preuves. Les fourchettes de la section 8 invitent à tester la robustesse ; elles ne justifient pas la confiance dans une date ou probabilité prédite de catastrophe.

Annexe A Principes conventionnels proposés

1. Les États restent responsables des systèmes d'IA développés, déployés ou substantiellement soutenus dans leur juridiction.
2. Les développeurs de pointe et fournisseurs de calcul désignés doivent enregistrer et déclarer les activités à haut risque aux autorités nationales compétentes.
3. L'entraînement et le déploiement au-delà des seuils convenus exigent un dossier de sécurité approuvé et une évaluation indépendante.
4. Les parties doivent protéger les poids de pointe et les infrastructures critiques selon des niveaux de sécurité communs.
5. Les incidents graves et quasi-incidents d'IA à potentiel transfrontalier ou catastrophique doivent être signalés par des canaux protégés.
6. Conserver l'autorité humaine licite d'emploi nucléaire et les vérifications indépendantes ; interdire la délégation à l'IA et la riposte automatique sans nouvelle décision humaine.
7. Les parties doivent maintenir un contrôle humain effectif sur les armes nucléaires et autres armes stratégiques et préserver les procédures de repli sans IA.
8. Les parties doivent mener des réexamens périodiques de sûreté nucléaire intégrant l'IA et fournir les attestations confidentielles convenues.
9. Utiliser déclarations protégées, preuves procédurales et droits d'inspection négociés, en reconnaissant l'incomplétude de la vérification et une procédure définie de réponse aux anomalies.
10. La non-conformité peut entraîner ordres correctifs, suspension de services de calcul ou modèles désignés, restrictions à l'exportation, sanctions financières et mesures collectives proportionnées.
11. Les informations de sécurité partagées avec l'agence internationale doivent bénéficier d'une forte protection juridique, technique et diplomatique.
12. Les seuils et annexes techniques doivent être réexaminés au moins annuellement afin de suivre les capacités plutôt que des mesures matérielles obsolètes.

Annexe B Dossier minimal de sécurité de l’IA de pointe

Section	Preuves requises
Gouvernance	Propriétaire juridique, dirigeant responsable, comité des risques, contestation indépendante, canal d’alerte, contact réglementaire.
Périmètre du système	Modèle, versions, outils, mémoire, données, identifiants, infrastructures, utilisateurs, juridictions et dépendances externes.
Modèle de menace	Abus, désalignement, initiés, vol de modèle, chaîne d’approvisionnement, compromission d’outils ou de moniteurs et défaillance corrélée.
Capacité	Évaluations standardisées et spécifiques, incertitudes, méthode de sollicitation des capacités, reproduction externe et prévision d’évolution à l’échelle.
Propension	Tromperie, détournement des récompenses, actions non autorisées, contournement de la supervision, comportement à l’arrêt et conscience de l’évaluation.
Contrôles	Isolement, identité, autorisation, journalisation, surveillance, limites de débit et de ressources, provenance, retour arrière et arrêt.
Risque résiduel	Fourchettes de probabilité et de gravité, hypothèses, avis experts dissidents, résultats adversariaux et anomalies non résolues.
Opérations	Enveloppe de déploiement, gestion des changements, seuils de surveillance, réponse aux incidents, rétablissement, retrait et plan de communication.
Assurance	Rapport d’évaluateur indépendant, conclusions du régulateur, preuves correctives et autorisation limitée dans le temps.

Annexe C Clauses types de lignes rouges nucléaire–IA

Article 1 Champ d’application

Un système d’IA est un système automatisé qui déduit, à partir d’entrées, comment générer des sorties telles que prédictions, contenus, recommandations ou décisions influençant des environnements physiques ou virtuels ; les systèmes varient en autonomie et en adaptabilité. Cette définition fonctionnelle s’inspire de l’article 3 du règlement européen sur l’IA.[5] Un système pertinent pour le nucléaire influence substantiellement l’alerte nucléaire, le renseignement, la préparation, les communications, le ciblage, la logistique ou l’aide à la décision. L’interdiction d’autorité vise la chaîne d’autorisation et d’exécution ; les applications auxiliaires restent soumises à des contrôles proportionnés. L’article 2 couvre aussi l’automatisation classique afin d’éviter une faille liée aux systèmes anciens.

Preuves de conformité : inventaire tenu à jour et décisions de classification documentées, accessibles à l’autorité habilitée désignée.

Article 2 Interdiction d’une autorité nucléaire autonome

Aucun État partie ne délègue à l’IA le pouvoir de décision d’emploi nucléaire, exclusivement ou conjointement, ni ne l’autorise à émettre, modifier ou authentifier un ordre d’emploi nucléaire ou à remplacer une autorisation humaine requise. L’emploi nucléaire exige une nouvelle décision de l’autorité humaine licite et des vérifications humaines positives et authentifiées indépendamment dans l’autorisation et l’exécution. L’initiation ou la riposte automatisées sans cette décision sont interdites, y compris les systèmes à base de règles. Des interverrouillages validés indépendamment peuvent empêcher une action dangereuse, mais non déclencher l’emploi nucléaire.

Preuves de conformité : architecture d’autorisation protégée, tests d’authentification humaine observés indépendamment et preuves d’exclusion de l’initiation automatisée classique.

Article 3 Contrôle humain et repli

Chaque État partie maintient des voies testées sans IA pour l’appréciation des alertes, les communications, l’authentification et l’aide à la décision. Les informations issues de l’IA sont identifiables et ne constituent pas, à elles seules, des preuves suffisantes pour l’emploi nucléaire. Les opérateurs peuvent isoler les composants d’IA par des contrôles indépendants en préservant les fonctions essentielles de sécurité. Interfaces et procédures permettent corroboration, désaccord et délibération validée pour la tâche. Les fenêtres d’escalade automatisées ne contournent pas la vérification humaine ; la mise en œuvre préserve les options sûres d’inhibition et n’impose pas de délais non testés.

Preuves de conformité : exercices d’isolement et de repli observés, évaluations des facteurs humains et dossiers de traitement des informations contradictoires.

Article 4 Réexamen de sûreté

Chaque État partie réalise un réexamen indépendant habilité intégrant l’IA à des intervalles n’excédant pas vingt-quatre mois et après tout changement substantiel ou incident grave. L’examen inclut les dépendances indirectes informationnelles et d’approvisionnement. Les exercices annuels sans IA testent des critères prédéfinis de restauration, continuité,

authentification et qualité décisionnelle. Les défaillances substantielles font l'objet de responsables correctifs nommés, d'échéances, de restrictions provisoires proportionnées et de nouveaux tests indépendants.

Preuves de conformité : dates d'examen, qualifications des examinateurs indépendants, résultats des exercices et clôture vérifiée des constats substantiels.

Article 5 Assurances négatives et vérification

Chaque État partie atteste annuellement sa conformité aux articles 2 à 4 et fournit les preuves protégées convenues. Les inspections internationales ordinaires concernent les installations civiles déclarées relevant de l'instrument. L'assurance classifiée combine examen national indépendant, rapports scellés et démonstrations encadrées. Les anomalies substantielles, dont attestations incohérentes, changements inexpliqués d'autorisations ou journaux obligatoires manquants, déclenchent consultation confidentielle et réponse dans un délai fixé. L'accès de contestation exige des garanties convenues. L'attestation seule ne doit pas être présentée comme une vérification exhaustive.

Preuves de conformité : déclarations signées, périmètre et limites de l'évaluateur, dossiers de garde et résolution documentée des anomalies ; protéger les secrets opérationnels.

Article 6 Incidents et communications de crise

Les parties maintiennent des contacts de crise authentifiés. Un risque crédible et imminent de fausse alerte nucléaire, d'action non autorisée ou d'escalade transfrontalière exige une notification immédiate par ce canal lorsque possible. Fournir une notification initiale sous vingt-quatre heures après identification, une première évaluation sous soixante-douze heures et des mises à jour continues ; viser un examen approfondi sous trente jours ou expliquer les investigations restantes. Les premiers rapports peuvent être incomplets et distinguent faits, hypothèses et confiance dans l'attribution. Ces délais sont proposés pour un traité ; ils ne décrivent pas un droit existant universel.

Preuves de conformité : contacts de crise exercés, notifications et corrections horodatées, enquêtes documentées distinguant incertitude et attribution.

Article 7 Confidentialité et conformité

Les informations échangées au titre de ces clauses bénéficient des protections juridiques et techniques convenues et ne sont pas utilisées pour obtenir un avantage commercial ou des renseignements militaires sans rapport. Les préoccupations de conformité sont d'abord soumises à une commission technique conjointe. Une non-conformité persistante peut entraîner consultations renforcées, demandes d'accès encadré, suspension de coopération ou mesures nationales proportionnées conformes au droit international.

Preuves de conformité : contrôles d'accès, règles de conservation et suppression, voie indépendante de réclamation et dossiers de décisions proportionnées susceptibles de réexamen.

Annexe D Sources et notes

Les noms des auteurs, les titres bibliographiques originaux et les liens sont conservés pour faciliter la vérification des sources.

[1] Bengio, Y. et al., International AI Safety Report 2026, DSIT 2026/001, 3 février 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[2] OpenAI, The Hugging Face incident and other third-party impact from misaligned models, mis à jour le septembre 2026, https://openai.com/hugging-face-incident-and-misalignment/

[3] OpenAI, Pacing model development in an era of cyber-critical capabilities, 18 août 2026, https://openai.com/index/pacing-model-development-cyber-capabilities/

[4] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework 1.0, janvier 2023, and Generative Artificial Intelligence Profile NIST AI 600-1, juillet 2024, https://www.nist.gov/itl/ai-risk-management-framework

[5] Regulation EU 2024/1689 laying down harmonised rules on artificial intelligence, Official Journal of the European Union, https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[6] Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, ouvert à la signature le 5 septembre 2024, https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[7] UK Department for Science, Innovation and Technology and Republic of Korea, Frontier AI Safety Commitments, AI Seoul Summit 2024, mis à jour le 7 février 2025, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

[8] Williams, I., AI Is Moving Into Nuclear Weapons Systems. Fail-Safe Reviews Are Critical, Nuclear Threat Initiative, 17 septembre 2026, https://www.nti.org/risky-business/ai-is-moving-into-nuclear-weapons-systems-fail-safe-reviews-are-critical/

[9] Hruby, J. and Miller, M. N., Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems, Nuclear Threat Initiative, 26 août 2021, https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/

[10] International Atomic Energy Agency, Computer Security for Nuclear Security, IAEA Nuclear Security Series No. 42-G, Vienna, 2021. https://www.iaea.org/publications/13629/computer-security-for-nuclear-security

[11] International Atomic Energy Agency, Computer Security Techniques for Nuclear Facilities, IAEA Nuclear Security Series No. 17-T Rev. 1, Vienna, 2021. https://www.iaea.org/publications/14729/computer-security-techniques-for-nuclear-facilities

[12] National Institute of Standards and Technology, Cybersecurity Framework 2.0, février 2024, https://www.nist.gov/cyberframework

[13] Cyberspace Administration of China and six other authorities, Interim Measures for the Management of Generative Artificial Intelligence Services, en vigueur le 15 août 2023, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[14] Cyberspace Administration of China and three other authorities, Measures for Labelling Artificial Intelligence Generated and Synthetic Content, en vigueur le 1 septembre 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

[15] Cyberspace Administration of China and other authorities, Interim Measures for the Management of Anthropomorphic Artificial Intelligence Interactive Services, 10 avril 2026, https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm

[16] Reuters, How China is preparing for the risk of AI escaping human control, 14 septembre 2026. Reportage sur les évolutions politiques ; les signatures et dates de révision ultérieures ne sont pas utilisées comme fondement. https://www.reuters.com/legal/litigation/how-china-is-preparing-risk-ai-escaping-human-control-2026-09-14/

[17] White House, America’s AI Action Plan, juillet 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[18] United Nations General Assembly, Resolution 79/325: Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance, 2025. https://docs.un.org/en/A/RES/79/325

[19] National Institute of Standards and Technology, Center for AI Standards and Innovation, model evaluations and standards work, consulté le 19 septembre 2026, https://www.nist.gov/caisi

[20] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report

[21] International Atomic Energy Agency, Safeguards and Verification, https://www.iaea.org/topics/safeguards-and-verification

[22] Financial Action Task Force, The FATF Recommendations, https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html

[23] OpenAI, The Hugging Face incident and the road ahead, 26 août 2026. Récit du développeur concernant l’incident de juillet. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

[24] METR and Redwood Research, Investigation of the OpenAI Hugging Face incident, 26 août 2026. Enquête externe avec une période limitée d’accès aux preuves. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[25] Anthropic, Investigating three real-world incidents in our cybersecurity evaluations, 30 juillet 2026. Divulgation du développeur. https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

[26] California, SB 53, Transparency in Frontier Artificial Intelligence Act, Chapter 138, approuvé le 29 septembre 2025. Texte législatif adopté. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[27] European Union, Regulation (EU) 2026/1744 (AI Omnibus); European Commission, AI Omnibus enters into force, 27 juillet 2026. Calendrier révisé d'application aux systèmes à haut risque et dispositions de supervision.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744

[28] White House, National Security Presidential Memorandum/NSPM-11, 5 juin 2026. Politique exécutive de sécurité nationale.
https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/

[29] White House, Readout of President Joe Biden's Meeting with President Xi Jinping of the People's Republic of China, 16 novembre 2024.
https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2024/11/16/readout-of-president-joe-bidens-meeting-with-president-xi-jinping-of-the-peoples-republic-of-china-3/

[30] Nuclear Threat Initiative, Advancing Nuclear Fail-Safe, présentation du projet décrivant l'examen américain achevé en 2024.
https://www.nti.org/about/programs-projects/project/advancing-nuclear-fail-safe/

[31] METR, Time Horizons, méthodologie et résultats ; voir aussi Kwa et al., Measuring AI Ability to Complete Long Tasks, arXiv:2503.14499 (2025).
https://metr.org/time-horizons/

[32] IETF, RFC 9334, Remote ATtestation procedureS (RATS) Architecture, janvier 2023.
https://www.rfc-editor.org/rfc/rfc9334.html

[33] National Telecommunications and Information Administration (2024). Dual-Use Foundation Models with Widely Available Model Weights Report. 30 juillet. Évaluation historique de politique publique, non un énoncé du droit universel actuel.
https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report

[34] Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P. and Awad, G. (2024). Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4.
https://doi.org/10.6028/NIST.AI.100-4

[35] Goddard, K., Roudsari, A. and Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. Publié en ligne le 16 juin 2011.
https://doi.org/10.1136/amiajnl-2011-000089

[36] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. and Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325. Recherche expérimentale ; ne constitue pas une validation de sûreté nucléaire.
https://arxiv.org/abs/2305.14325

[37] Mosier, K. L. and Skitka, L. J. (1999). Automation Use and Automation Bias. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43(3), 344–348.
https://doi.org/10.1177/154193129904300346

Base probante et limites

Date limite des éléments examinés : 20 septembre 2026. Les textes officiels n'établissent des obligations que dans leur juridiction et selon leurs dates d'application. Les divulgations de développeurs sont des récits primaires de parties intéressées ; les enquêtes externes améliorent l'examen tout en restant limitées par l'accès et la sélection. Les évaluations techniques établissent des performances dans les conditions testées. Les commentaires de politique publique apportent des arguments, non des estimations de fréquence. Une confiance élevée indique un fort soutien convergent ; modérée, un soutien crédible mais incomplet ; faible, un soutien fragmenté ou sensible aux hypothèses ; très faible, de fortes limitations. Ces labels sont distincts de la probabilité. Aucune fourchette de la section 8 n'a été calibrée indépendamment. Il s'agit d'une étude de politique publique signée, non d'une prédiction validée expérimentalement ni d'un article de revue formellement évalué par les pairs.

Annexe E Protocole de mise en œuvre et d'assurance

Cette annexe propose un langage de politique publique pour adapter l'étude à une réglementation nationale, une directive de sécurité nationale ou un instrument conventionnel négocié. Elle ne crée aucune obligation actuelle. Ses exigences visent les activités couvertes de niveau 3 et les applications permises de niveau 4B ; l'autorité d'emploi nucléaire interdite de niveau 4A ne peut être autorisée par un dossier de sécurité. Les normes sont neutres quant à l'origine et s'appliquent aux systèmes ouverts ou fermés selon capacité, accès et conséquences.

E.1 Responsabilités et modifications substantielles

Développeur initial. Avant une diffusion couverte, documenter l'identité du modèle, le périmètre d'évaluation, les limites connues, les améliorations plausibles en aval et l'irréversibilité de la distribution. Fournir aux régulateurs et destinataires légitimes des informations de sécurité proportionnées et un contact d'incident. Protéger les détails exploitables par divulgation contrôlée. La décision de diffusion devrait préciser les usages évalués et ceux restant hors du champ probant.

Modificateur substantiel ou auteur d'ajustement fin. Consigner le modèle parent lorsqu'il est connu, les identifiants de version, les changements pertinents et les preuves d'évaluation obtenues. Une modification substantielle comprend une hausse démontrée de capacité dangereuse, la suppression d'un contrôle essentiel, l'extension de l'autonomie ou l'adaptation à un usage plus lourd de conséquences. Soumettre un dossier modifié ou nouveau avant diffusion ou déploiement couverts. Les changements à faible impact dans l'enveloppe approuvée exigent des vérifications de non-régression documentées, sans déclencher automatiquement une approbation complète.

Intégrateur et opérateur. Identifier la personne morale responsable du système déployé. Évaluer la combinaison du modèle, des sources documentaires, des outils, de la mémoire, des identifiants, de la surveillance et des procédures humaines. L'approbation antérieure d'un modèle n'autorise pas de nouvelles permissions ni une application nucléaire. Tenir un registre des changements, nommer un responsable des incidents, tester l'isolement et maintenir l'enveloppe d'exploitation à jour.

Distributeur et fournisseur d'infrastructure. Lorsque la législation couvre leur activité, conserver les dossiers minimaux permettant d'identifier opérateurs réglementés et autorisations, préserver les preuves pertinentes et répondre aux ordres légaux susceptibles de réexamen. Ne pas imposer une surveillance universelle des utilisateurs ordinaires ni supposer que tout dépôt passif peut évaluer chaque modèle téléversé. Les obligations doivent être proportionnées au contrôle réel, à l'échelle, à la connaissance et au risque.

Responsabilité et preuves. Le régulateur national devrait inspecter évaluations de diffusion, déclarations de filiation, journaux de changements, rapports d'évaluation et autorisations d'exploitation. Attribuer la responsabilité selon l'obligation violée, la causalité, la prévisibilité et le contrôle de l'acteur, avec les protections procédurales applicables. Les contrats peuvent répartir le travail, non effacer les obligations légales. Une filiation absente ou trompeuse est une déficience d'assurance ; une empreinte de modèle est un identifiant, non une preuve de sécurité ou d'ascendance exhaustive.

E.2 Application transfrontalière et IFASA

Un modèle hébergé à l'étranger ou téléchargé d'un laboratoire étranger reste soumis aux règles applicables de l'État importateur ou déployeur lorsqu'il sert une activité intérieure couverte. Les États participants devraient désigner une autorité, ne reconnaître les évaluations que dans leur périmètre testé et échanger des notifications protégées de défaillances graves. L'IFASA devrait soutenir méthodes communes, accréditation, coordination des incidents et consultation technique. Les pouvoirs d'application doivent découler du droit interne et du mandat conventionnel convenu.

Pour un opérateur couvert, une déficience étayée devrait permettre à l'autorité d'imposer une réévaluation, de limiter les autorisations, de suspendre le déploiement affecté ou de sanctionner proportionnellement. Les restrictions d'urgence exigent des motifs documentés et un réexamen rapide. Les constats concernant un développeur étranger devraient distinguer preuves vérifiées et allégations, avec procédure de correction ou contestation. La nationalité seule ne prouve pas la non-conformité.

Pour les poids copiés hors des juridictions participantes, les mesures restent nécessairement incomplètes : coopération volontaire, restrictions des usages et services nationaux réglementés, échange protégé sur les incidents et mesures légalement autorisées contre les conduites nuisibles identifiables. Aucun avis, condition de licence ou enregistrement ne garantit l'effacement des poids distribués. Signaler la lacune résiduelle d'application plutôt que prétendre à un retrait mondial ou à un pouvoir international d'arrêt.

E.3 Intégrité de l'information en crise

Acteur responsable et déclencheur. L'autorité nucléaire responsable applique une procédure documentée d'intégrité lorsque des informations synthétiques, contestées ou résumées par IA peuvent affecter substantiellement l'appréciation de crise ou la posture nucléaire. Un responsable de permanence désigné conserve la responsabilité de la vérification ; un classificateur d'IA n'est pas l'arbitre final de l'authenticité.

Contrôles requis. Identifier source, date, historique de traitement disponible, statut d'authentification et contradictions non résolues. Séparer faits observés, allégations et inférences. Utiliser des corroborations de sources indépendantes et des canaux officiels authentifiés. La répétition entre médias ou l'accord de modèles utilisant la même entrée ne

constitue pas une confirmation indépendante. Préserver les originaux et éviter les synthèses effaçant l'incertitude. Une origine authentique ne garantit pas l'exactitude factuelle ; l'absence de provenance ne prouve pas la falsification.

Preuves d'assurance. Dans des exercices protégés, tester contenus fabriqués et authentiques, usurpation, provenance indisponible, sources de confiance compromises et communications dégradées. Consigner fausses acceptations et rejets, délai de vérification, contradictions manquées et erreurs décisionnelles. Prédéfinir les performances acceptables pour chaque fonction et condition de crise dans le dossier sectoriel ; ne pas inventer un délai ou seuil d'exactitude universel. Tester face à une vérité d'exercice établie indépendamment, non à un corrigé généré par modèle.

Échec et réponse. L'échec d'un critère essentiel impose des restrictions proportionnées sur la voie informationnelle assistée concernée, la restauration de solutions validées, des corrections et une nouvelle épreuve observée. Préserver l'alerte et l'inhibition sûre. Aucune vidéo non authentifiée, score de média synthétique ou synthèse d'IA ne satisfait seul un préalable à l'emploi nucléaire. Cette garantie ne justifie pas d'ignorer d'autres preuves corroborées indépendamment.

E.4 Contrôle humain effectif et contestation indépendante

Fonction requise. Les appréciations assistées par IA à conséquences majeures permettent une contestation fondée sur un raisonnement indépendant, l'accès aux preuves et une présentation claire des désaccords non résolus. Les humains qualifiés conservent autorité et responsabilité. Lorsque le temps et la sécurité le permettent, consigner une appréciation humaine initiale avant de révéler les recommandations d'IA. Sinon, utiliser des procédures validées pour la tâche ; ne pas imposer de délais arbitraires empêchant l'action protectrice.

Débat entre IA comme mécanisme facultatif. Un second modèle ne peut soutenir la contestation que dans un dossier approuvé. Documenter les dépendances communes connues de formation, données, recherche documentaire, outils et surveillance. Obtenir les appréciations initiales indépendamment avant tout échange, préserver les dissensions et tester si le débat corrige ou propage les erreurs. L'accord de modèles n'est pas une preuve indépendante ; un vote d'IA ne détermine pas l'autorité d'emploi nucléaire. Des fournisseurs différents ne garantissent pas des modes de défaillance indépendants.

Validation. Comparer dispositifs humains seuls, à conseiller unique et à plusieurs conseillers proposés dans des exercices représentatifs protégés. Inclure entrées trompeuses, modèles conjointement erronés, explications convaincantes non étayées, preuves contradictoires, indisponibilité de l'IA et pression temporelle. Préenregistrer les critères de réussite et consigner l'incertitude. Mesurer fausse acceptation, dangers manqués, qualité décisionnelle, temps de vérification, charge et rétablissement après retrait de l'IA. Évaluer tout le système opérateur–interface, non la seule exactitude aux benchmarks.

Condition de déploiement et réexamen. L'autorité habilitée ne devrait permettre une configuration de contestation par IA que si les preuves soutiennent un bénéfice de sécurité sans délai ni charge inacceptables. Surveiller les dérives et répéter la validation après changements substantiels, lors des réexamens périodiques et après incidents graves. Suspendre ou restreindre toute configuration échouant à un critère essentiel. Maintenir une voie de contestation sans IA. Ce sont des exigences d'assurance proposées ; les études citées ne fixent pas de seuil universel de performance nucléaire.[35–37]

E.5 Notification et autorisation du calcul

Acteurs et déclencheurs couverts. Les développeurs et fournisseurs de services désignés d'entraînement à grande échelle devraient déclarer le calcul couvert selon des règles publiées. Les déclencheurs de capacité et de déploiement restent applicables indépendamment, y compris aux ajustements fins peu coûteux et aux modèles importés. Un montant tel que 100 millions USD ne constitue ni une exemption de sécurité sous ce montant ni une preuve concluante de danger au-dessus.

Déclarations et agrégation. Consigner entité responsable, programme coordonné, fournisseurs participants, estimations de calcul et incertitudes, post-entraînement substantiel, autorisation et étapes d'évaluation. Définir l'agrégation des entraînements coordonnés et de la sous-traitance entre lieux et entités. Auditer indépendamment un échantillon fondé sur le risque et enquêter sur les écarts. Ni propriété du matériel ni structure de facturation ne doivent permettre à des activités réglementées équivalentes d'échapper à la déclaration.

Application et réexamen. Dissimulation ou contournement substantiels devraient déclencher enquête et mesures légales proportionnées. Protéger les informations confidentielles et permettre la correction des erreurs d'estimation de bonne foi. Réexaminer annuellement au moins les seuils selon capacités observées, efficacité algorithmique et expérience d'application, avec motifs et dispositions transitoires. Les dossiers de calcul soutiennent l'assurance ; ils ne prouvent pas la détection de tout modèle non déclaré ou opération clandestine.

E.6 Adoption par directive ou traité

Une directive de sécurité nationale devrait identifier autorité émettrice et base juridique, agences et prestataires couverts, dirigeant responsable, financement, canal de rapport et organe indépendant habilité. Séquence indicative : attribuer les responsabilités et inventorier les systèmes sous quatre-vingt-dix jours ; soumettre les écarts et garanties provisoires sous cent quatre-vingts jours ; achever les premiers exercices représentatifs et plans correctifs sous douze mois. Ce sont des étapes proposées, sous réserve de l'autorité légale et des exigences opérationnelles de la juridiction.

Une annexe conventionnelle devrait identifier obligations des États parties, périmètre du consentement, relations entre IFASA et autorités nationales, formats de preuves protégées, intervalles d'examen, droits de consultation et procédures de différend. Préciser comment l'application interne atteint les acteurs privés pertinents et comment l'accès encadré est autorisé. Aucune clause ne devrait impliquer l'inspection illimitée des systèmes classifiés ni une autorité juridique sur les non-parties. Ratification, entrée en vigueur et application exigent une négociation distincte.

Dossier de suivi. Chaque autorité participante devrait conserver un registre protégé identifiant acteur responsable, activité déclenchante, preuves soumises, décision, limites résiduelles, échéance corrective et prochain examen. Publier des résultats agrégés lorsque cela est sûr. Les examinateurs indépendants devraient tester si les contrôles modifient réellement les pratiques plutôt que produire seulement des déclarations. L'annexe soutient une application auditable ; elle ne transforme pas des preuves incomplètes en démonstration de conformité universelle.