

PROPOSITION 46

Soumettre les intelligences artificielles critiques à une supervision internationale préalable

KAIROS | Fiche de synthèse révisée | Repère de calendrier : année n

DOCTRINE — Plus le dommage potentiel est grave, transnational et irréversible, plus le contrôle doit intervenir en amont.

Pourquoi agir ?

Les intelligences artificielles ne présentent pas toutes les mêmes dangers. Un assistant de rédaction et un système capable de faciliter une attaque contre un réseau électrique ne doivent pas être soumis aux mêmes contrôles.

- En 2022, la diffusion publique de ChatGPT a accéléré l'adoption de l'IA générative. La vitesse de déploiement peut dépasser celle de la préparation des autorités chargées de la sécurité.
- En 2024, l'Union européenne a adopté son règlement sur l'IA. Il distingue déjà plusieurs catégories de risques, mais une technologie développée dans un pays peut être utilisée immédiatement dans plusieurs autres.
- Une cyberattaque majeure contre des hôpitaux, des réseaux électriques ou des systèmes financiers peut produire des effets au-delà des frontières. L'IA pourrait amplifier certaines capacités d'attaque ; cela justifie des tests rigoureux, sans présumer que chaque modèle est dangereux.
- Les risques biologiques extrêmes, l'autonomie incontrôlée et le contournement des protections exigent des évaluations scientifiques indépendantes. Les scénarios catastrophiques restent incertains : il faut les examiner, pas les présenter comme des événements déjà survenus.
- Les entreprises ont intérêt à démontrer leur fiabilité. Les citoyens, les chercheurs et les États doivent pouvoir connaître les garanties essentielles sans accéder aux secrets industriels.

L'enjeu est donc ciblé : vérifier en amont les très rares systèmes présentant un potentiel de dommage grave, transnational ou difficilement réversible. Cette exigence ne doit pas freiner la recherche ordinaire ni transformer toute innovation en activité soumise à autorisation internationale.

Ce que nous apprend l'Histoire

L'histoire des technologies sensibles montre qu'une coopération internationale peut rendre les risques plus visibles et les contrôles plus crédibles. Elle rappelle aussi qu'aucun modèle institutionnel ne se transpose automatiquement à l'intelligence artificielle.

- En 1957, l'Agence internationale de l'énergie atomique (AIEA) est créée. Elle développe des mécanismes de vérification et d'inspection dans le domaine nucléaire. L'IA n'est pas une matière nucléaire : l'exemple porte sur la méthode de contrôle, non sur une identité des risques.
- En 1944, la convention de Chicago pose les bases de l'aviation civile internationale. Des normes communes facilitent ensuite la sécurité d'activités traversant quotidiennement les frontières.
- En 1986, l'accident de Tchernobyl révèle les conséquences transnationales d'un événement technologique majeur. Les conventions internationales de notification et d'assistance nucléaires sont adoptées la même année.
- En 1997, la Convention sur les armes chimiques entre en vigueur et s'appuie sur une organisation de vérification. Elle illustre l'importance d'inspections encadrées par des obligations juridiques précises.
- En 2023 et 2024, plusieurs sommets internationaux consacrés à la sûreté de l'IA soulignent la nécessité d'évaluations et d'échanges entre gouvernements, chercheurs et entreprises.

Ces précédents enseignent trois choses : définir exactement le risque contrôlé, confier la vérification à des experts indépendants et prévoir des voies de recours. Ils montrent également que la confiance internationale se construit

progressivement. Un traité crédible peut commencer avec une coalition d'États volontaires plutôt qu'attendre un accord universel.

Ce qui existe aujourd'hui

Plusieurs règles encadrent déjà l'intelligence artificielle. La proposition KAIROS ne les remplace pas : elle cherche à combler une lacune éventuelle concernant les systèmes les plus dangereux à l'échelle internationale.

- Le règlement européen sur l'intelligence artificielle, adopté en 2024, prévoit une approche graduée. Certaines pratiques sont interdites ; des obligations spécifiques concernent les systèmes à haut risque et les modèles à usage général présentant un risque systémique.
- Le règlement s'applique par étapes. Son calendrier doit être vérifié lors de la mise en œuvre, car certaines dispositions peuvent évoluer et les obligations ne commencent pas toutes le même jour.
- La convention-cadre du Conseil de l'Europe sur l'IA, ouverte à la signature en 2024, traite des droits humains, de la démocratie et de l'État de droit. Elle n'instaure pas à elle seule une agence mondiale de certification technique.
- Les Nations unies développent des espaces de dialogue et d'expertise scientifique sur l'IA. Ces démarches favorisent la coopération, mais ne constituent pas une autorisation internationale obligatoire avant le déploiement d'un modèle critique.
- Des laboratoires publics et privés réalisent déjà des tests de sécurité, notamment des exercices de « red team » : des équipes tentent de détecter les failles avant qu'elles ne soient exploitées.

Le point faible reste la comparabilité des tests, leur indépendance et la reconnaissance des résultats entre pays. Il faut éviter de créer des obligations redondantes : une certification internationale ciblée devrait compléter les contrôles nationaux et européens, protéger les données sensibles et réserver les procédures les plus lourdes aux risques réellement exceptionnels.

Ce que nous proposons

KAIROS propose une Agence internationale de vérification et de sûreté de l'intelligence artificielle (AIVIA), créée par traité. Elle serait réservée aux « IA critiques internationales » : une catégorie étroite, définie selon des capacités démontrées et des dommages potentiellement catastrophiques ou transnationaux.

- Définir les seuils : expertise scientifique indépendante sur les risques biologiques extrêmes, les cyberattaques sophistiquées, l'autonomie incontrôlée et les contournements graves des protections. Les seuils seraient régulièrement révisés.
- Déclarer en amont : lorsqu'un développeur prévoit raisonnablement d'atteindre le seuil critique, il notifie confidentiellement le projet. Des exigences particulières peuvent alors encadrer l'entraînement et les accès.
- Tester sans divulguer : laboratoires accrédités, équipes de « red team », inspections sécurisées et contre-évaluations. Les secrets industriels et les informations exploitables pour nuire restent protégés.
- Certifier avant d'autoriser : l'AIVIA délivre, conditionne ou refuse un certificat. L'État ou l'Union conserve la décision juridique de déploiement, mais les signataires du traité ne pourraient autoriser un système critique sans certificat valide.
- Surveiller après déploiement : déclaration des incidents graves, nouvelles évaluations, restrictions d'usage et procédures crédibles de retrait ou d'arrêt. Une alerte exceptionnelle serait motivée, temporaire et susceptible de recours.
- Préserver la proportionnalité : l'IA ordinaire, la recherche non critique et les petits modèles sans capacités dangereuses restent hors de ce régime. Les usages militaires feraient l'objet de négociations distinctes.

Le principe est simple : plus un dommage possible est grave et difficile à réparer, plus le contrôle doit intervenir tôt, sous responsabilité publique et avec des garanties vérifiables.

Calendrier de mise en œuvre

Le repère « année n » désigne l'année de lancement politique de la proposition, sans présumer de sa date civile. L'entrée en vigueur internationale dépendra des négociations et des ratifications : les étapes ci-dessous sont donc des objectifs, pas des échéances garanties.

- Année n : la France propose une doctrine européenne commune. Une coalition de pays volontaires définit les risques critiques, les principes de confidentialité et un mandat de négociation. Publication d'une première étude juridique et budgétaire.
- Année n+1 : mise en place d'un protocole volontaire. Des laboratoires participants harmonisent leurs méthodes de test ; un registre confidentiel des incidents graves et un réseau de chercheurs indépendants sont expérimentés.
- Année n+2 : premiers exercices d'évaluation croisée, simulations d'inspection et certificats pilotes sans valeur obligatoire. Un audit vérifie les coûts, la sécurité des informations et les effets sur l'innovation.
- Années n+2 à n+3 : négociation du traité : définition des seuils, pouvoirs d'inspection, conditions de certification, procédures d'urgence, droits de recours et articulation avec le droit européen.
- Année n+4 : si un nombre suffisant d'États ratifie, création opérationnelle de l'AIVIA, recrutement d'experts et accréditation des premiers laboratoires. Sinon, poursuite de la coopération volontaire.
- Années n+5 à n+6 : premières certifications et inspections obligatoires dans les États parties, uniquement pour les systèmes entrant dans la catégorie critique.
- Année n+7 : bilan scientifique, financier et démocratique indépendant ; révision des seuils et des méthodes. Tous les deux ans ensuite, publication d'un rapport de performance et adaptation des protocoles aux progrès techniques.

Combien ça coûte ? Comment on finance ?

La fiche source définit les financeurs possibles, mais ne chiffre pas l'AIVIA. KAIROS propose donc ici un scénario indicatif à confirmer par une étude d'impact, selon le nombre d'États signataires, de laboratoires et de systèmes effectivement classés critiques.

- Phase préparatoire, années n à n+3 : 24 millions d'euros au total, soit 6 millions par an en moyenne. Répartition proposée : 8 millions pour les méthodes et laboratoires pilotes ; 6 millions pour les évaluations croisées ; 4 millions pour la sécurité des données ; 4 millions pour la négociation et le cadre juridique ; 2 millions pour l'évaluation indépendante.
- Financement de cette préparation : 12 millions de contributions publiques des États volontaires, 8 millions de programmes européens de recherche et de coopération, et 4 millions de redevances ou participations des grands opérateurs pilotes, sans droit de regard sur les résultats.
- Fonctionnement de l'AIVIA après traité : hypothèse centrale de 60 millions d'euros par an. Elle comprend 22 millions pour les experts et inspections, 14 millions pour les laboratoires et tests, 10 millions pour la cybersécurité, 6 millions pour les registres et l'alerte, et 8 millions pour l'assistance technique et la gouvernance.
- Ressources annuelles envisagées : 36 millions de contributions obligatoires des États parties, 18 millions de redevances proportionnées sur les opérateurs critiques et 6 millions de programmes internationaux ou européens dédiés.

Ces sommes ne sont ni votées ni garanties. Les petits acteurs ne financeraient pas le coût principal. Un audit contrôlerait les comptes, les conflits d'intérêts et les certifications ; le budget serait révisé après deux exercices.

Note de méthode : les chiffres budgétaires sont des hypothèses de travail ; ils ne figurent pas dans la fiche source et ne constituent pas des crédits acquis. Les exemples historiques sont des comparaisons, non des preuves d'équivalence des risques.