Soumettre les intelligences artificielles critiques à une supervision internationale préalable

La proposition en clair

Conditionner le déploiement des IA à risque critique à une évaluation préalable et promouvoir une supervision internationale vérifiable.

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.

Situation actuelle

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.

Mesures proposées

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.

Coûts et financement

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.

Source : page KAIROS « proposition-46.html » transmise le 29 octobre 2026. Cette fiche résume les éléments publiés ; consulter la page pour le texte complet.