Vers une expertise française en évaluation de l'IA : Le rôle stratégique de l'INESIA

Centre pour la Sécurité de l'IA

Introduction

C'est à la veille du Sommet pour l'Action sur l'IA que la France a annoncé la mise en place de son Institut National pour l'Évaluation et la Sécurité de l'Intelligence Artificielle (INESIA)[1]. Établi en application des engagements de la déclaration de Séoul de mai 2024[2], cet institut national voit le jour à un moment charnière pour l'encadrement de l'IA. Alors que la compétition internationale bat son plein, que les alertes de la communauté scientifique se font de plus en plus pressantes et que le règlement européen sur l'IA entre pleinement en vigueur, l'INESIA doit contribuer à doter la France de capacités de pointe d'évaluation des systèmes d'intelligence artificielle. L'objectif est double : d'une part, assurer un encadrement efficace des systèmes d'IA avancés dont le déploiement s'accélère en France et en Europe, et d'autre part, renforcer l'influence française dans les débats internationaux sur le sujet.

Suite aux annonces concernant le mandat et la structure de l'INESIA, la priorité est de définir précisément ses axes de travail.

Auteurs: Jérôme Barbier, Charles Martinet, Charbel-Raphaël Segerie

[1] https://www.economie.gouv.fr/actualites/la-france-se-dote-dun-institut-national-pour-levaluation-et-la-securite-de-lintelligence
[2] Voir la déclaration d'intention et la déclaration ministérielle: https://www.industry.gov.au/publications/seoul-declaration-countries-attending-ai-seoul-summit-21-22-may-2024#seoul-declaration-1.

I - Un institut pour la sécurité de l'IA à la française

S'inscrivant dans une dynamique internationale visant à promouvoir une IA sûre, novatrice et inclusive, l'établissement de l'INESIA est aussi l'occasion d'affirmer un modèle français pour l'évaluation des systèmes d'intelligence artificielle. L'institut est chargé par l'État de conduire trois missions principales : le soutien à la mise en œuvre de la régulation de l'IA en application notamment du nouveau règlement européen ; l'évaluation de la performance et de la fiabilité des modèles d'IA (la mission centrale de l'ensemble des instituts nationaux pour la sécurité de l'IA établis dans le cadre du processus initié à Bletchley park) ; mais aussi l'analyse des risques systémiques liés à l'IA dans le champ de la sécurité nationale[3].

Contrairement à la plupart des instituts homologues, qui prennent la forme d'une organisation nouvelle ou sont confiée à une entité unique existante (agence, direction ministérielle, régulateur, mais aussi université ou organisation de la société civile), l'INESIA se présente comme un consortium d'opérateurs de l'État, de directions ministérielles et de services interministériels. Ce consortium, qui ne donne lieu à aucune nouvelle création juridique, regroupe l'Agence nationale de la sécurité des systèmes d'information (ANSSI), l'Institut national de recherche en sciences et technologies du numérique (Inria), le Laboratoire national de métrologie et d'essais (LNE) et le Pôle d'expertise de la régulation numérique (PEReN). Le pilotage de l'institut sera conjointement assuré par le Secrétariat Général pour la Défense et la Sécurité Nationale (SGDSN) et la Direction Générale des Entreprises (DGE) du Ministère de l'Economie et des Finances.

Cette structuration présente plusieurs avantages. Elle permet une collaboration étroite entre l'ensemble des acteurs de l'État compétents sur l'une ou l'autre composante de l'évaluation et de la sécurité de l'IA, permettant de s'appuyer sur les compétences existantes. De plus, l'INESIA est le seul institut de ce type disposant d'un mandat explicite sur les enjeux de sécurité nationale, ce qui lui confère une compétence unique pour appréhender le continuum sûreté-sécurité-défense – clé pour cette technologie d'usage général. Ce choix offre à la France des opportunités d'influence significatives sur la scène internationale, en lui permettant de jouer un rôle actif au sein du réseau international des instituts nationaux sur la sécurité de l'IA, aux côtés de nations telles que le Canada, la Corée du Sud, les États-Unis, le Japon, le Kenya, Singapour et le Royaume-Uni.

Ce choix présente cependant plusieurs risques en comparaison des décisions prises par plusieurs partenaires de la France. L'absence d'une structure juridique unique et d'une tutelle unifiée pour l'INESIA pose notamment la question de la pérennité de son financement et de ses ressources humaines, en comparaison d'instituts étrangers aux moyens souvent votés sur une base pluriannuelle. L'institut britannique, fondé comme une "start up au sein du gouvernement" s'est par exemple vu doté de 100 millions de livres sterling jusqu'en 2030 - une clarté budgétaire qui permet à l'institut de développer en toute sérénité une stratégie sur les moyen et long termes.

[3] Cf. l'annonce du mandat dans le communiqué de presse du ministère de l'Economie et des Finances: https://www.economie.gouv.fr/actualites/la-france-se-dote-dun-institut-national-pour-levaluation-et-la-securite-de-lintelligence

De même, la largesse du mandat et la diversité des entités publiques impliquées dans le consortium présentent un risque de diffusion de l'action de l'Etat en matière d'évaluation et de sécurité des systèmes et modèles d'intelligence artificielle. Un effort de fond devra être entrepris pour constituer une culture de travail commune à l'ensemble du consortium et à prévenir le travail en silo sur les différentes priorités de l'INESIA. Parallèlement et afin de prévenir les risques de rivalités institutionnelles inhérents à l'absence d'une hiérarchie fonctionnelle claire au sein du consortium, il est impératif de structurer la gouvernance opérationnelle de l'INESIA. Il convient d'instaurer un mode de pilotage fondé sur la spécialisation, où l'entité membre disposant de l'expertise reconnue prend l'initiative sur chaque sujet spécifique.

Caractéristiques des principaux instituts nationaux pour la sécurité de l'IA[4]

Pays (Organisation)	Date de création	Activités principales	Institution d'accueil	Financement
France (INESIA)	3 février 2025 (annonce)	- Analyse des risques systémiques dans le champ de la sécurité nationale - Soutien à la mise en œuvre de la régulation de l'IA - Evaluation de la performance et de la fiabilité des modèles d'IA.	Consortium d'entités publiques	Non communiqué - pas de financement ad hoc
Singapour (Digital Trust Center)	1er juin 2022	- Test et évaluation - Conception, développement et déploiement de modèles sécurisés - Traçabilité des contenus	- Info-communications Media and Development Authority (agence publique) - AI Verify Foundation (ONG)	37 million $ au lancement en 2022
Royaume-Uni (UK AISI)	24 avril 2023	- Evaluation (pré- et post-déploiement) - Recherche fondamentale en sécurité de l'IA - Publication du rapport "état de la science"	Department of Science, Innovation and Technology (ministère de plein exercice)	100 millions £ (sécurisé jusqu'en 2030)

[4] Araujo, R., et al. Understanding the First Wave of AI Safety Institutes. Institute for AI Policy & Strategy. https://www.iaps.ai/research/understanding-aisis; Reddel, M., et al. The AI Safety Institute Network: Who, What and How? Center for Future Generations. https://cfg.eu/the-ai-safety-institute-network-who-what-and-how/

Pays (Organisation)	Date de création	Activités principales	Institution d'accueil	Financement
Etats-Unis (US AISI)	2 février 2024	- Evaluation (pré- et post-déploiement) - Recherche fondamentale en sécurité de l'IA (détection de deepfakes, sécurité des modèles, bonnes pratiques, standards) - Organisation de la rencontre du réseau des instituts pour la sécurité de l'IA (San Francisco, novembre 2023)	National Institute for Standards and Technology, agence indépendante au sein du Département au Commerce	10 millions $ pour 2024 et 2025
Japon (Japan AISI)	4 février 2024	- Evaluation des modèles (conception et mise en oeuvre) - Recherche sur de possibles standards - Coordination internationale	Information-technology Promotion Agency (agence publique)	Peu clair
Union Européenne (EU AI Office)	29 mai 2024	- Standardisation et développement de codes de conduite (recherche pour un alignement international) - Mise en oeuvre et sanction de l'AI Act - Evaluation des risques systémiques liés aux modèles	Commission Européenne (DG-CONNECT)	46,5 millions € au lancement

Centre pour la Sécurité de l'IA, 2025.

Pays (Organisation)	Date de création	Activités principales	Institution d'accueil	Financement
Canada (Canadian AISI)	12 novembre 2024	- Évaluation et tests des systèmes d'IA - Recherche fondamentale en sécurité de l'IA - Développement de recommandations pour l'atténuation des risques - Coordination avec la communauté canadienne de recherche et les entreprises - Collaboration internationale avec d'autres instituts	Ministère de l'Innovation, Sciences et Développement économique	50 millions $ sur 5 ans.
Corée du Sud (Korea AISI)	27 novembre 2024	- Recherche sur la sécurité de l'IA et méthodes préemptives d'identification des risques - Recherche sur de possibles standards - Coordination internationale	Electronics and Telecommunications Research Institute (ONG soutenue par fonds publics)	Non communiqué
Inde (India AISI)	31 janvier 2025	- Recherche sur la sécurité de l'IA - Méthodes de test pour les modèles - Développement d'outils d'authentification des contenus (*watermarking* et *labelling*)	IndiaAI (initiative publique-privée portée par le Ministère de l'Electronique et des Technologies de l'Information)	64 millions de dollars (5,51 milliards de roupies) pour la mission IndiaAI pour 2025[5]

[5] Le budget exact consacré à la priorité sécurité n'a pas été publié

Centre pour la Sécurité de l'IA, 2025.

II - Développer des capacités souveraines de pointe en matière d'évaluation des modèles

___

Les instituts nationaux pour la sécurité de l'IA ont été pensés lors des Sommets IA comme ponts entre la gouvernance politique et l'expertise technique nécessaire pour évaluer les modèles d'IA avancés. Par conséquent, il est peu étonnant que la légitimité de ces instituts se soit avant tout fondée sur leur capacité à développer une forte expertise technique en interne. L'INESIA doit s'inspirer de l'expérience des instituts homologues afin de fonder sa légitimité sur une expertise technique approfondie, dont dépendra avant tout autre chose sa crédibilité à l'échelle nationale, européenne, et internationale. Pour développer cette expertise, nous conseillons de structurer en trois étapes l'initiation des travaux de l'INESIA :

- Premièrement, l'INESIA doit commencer par établir un état des lieux des méthodologies d'évaluation existantes, suivi d'une analyse détaillée de leurs forces et de leurs lacunes respectives. Cela l'aidera à choisir entre l'utilisation d'outils et méthodes existants et le développement de nouveaux outils.
- Dans un second temps, l'INESIA pourra consacrer ses premiers travaux à la reproduction des évaluations déjà réalisées et décrites par ses homologues, permettant de vérifier leur robustesse tout en développant une expertise pratique et en identifiant des axes d'amélioration pour le développement de ses propres outils et méthodes d'évaluation.
- Ce travail préliminaire doit rapidement déboucher sur le travail de fond : l'évaluation proprement dite des modèles déjà ou prochainement déployés. Le défi principal réside dans la conception des protocoles d'évaluation suffisamment robustes pour analyser des systèmes d'IA extrêmement puissants, polyvalents et peu voire pas interprétables.

L'INESIA devra affronter trois types de difficulté :

1. **L'évaluation doit dépasser le modèle lui-même.**

   La difficulté majeure réside dans le fait que les risques ne proviennent pas seulement du modèle algorithmique isolé, mais de l'ensemble de ses usages potentiels et réels. Il est crucial d'évaluer non seulement le modèle, mais aussi, lorsqu'il existe, le système complet, incluant son interface et son contexte d'utilisation. Plus particulièrement, l'évaluation doit intégrer une analyse approfondie de la capacité potentielle du modèle à interagir avec le monde réel et virtuel, par exemple son aptitude à utiliser des API ou à naviguer sur internet. Une évaluation complète doit également considérer non seulement les capacités intrinsèques du modèle (ce qu'il peut faire), mais aussi ses tendances comportementales (ce qu'il est enclin à faire) et les impacts à grande échelle sur la société et l'économie qui peuvent en découler.

2. **Les méthodes d'évaluation actuelles sont insuffisantes.**

   Les outils et méthodologies disponibles pour évaluer les modèles d'IA sont imparfaits et peinent à suivre le rythme effréné du développement technologique. Ces méthodes souffrent de problèmes de conception, d'une couverture incomplète des risques (notamment pour les systèmes multimodaux, les interactions humaines et les impacts systémiques) et de difficultés d'interprétation des résultats[6].

3. **L'opacité intrinsèque des modèles complique leur évaluation.**

   Les modèles d'IA sont des "boîtes noires" dont le fonctionnement interne est difficilement interprétable. Cette opacité pose un défi fondamental, car les modèles peuvent développer des comportements imprévus, voire délibérément trompeurs. Ils peuvent, par exemple, dissimuler certaines capacités dangereuses lors des tests ("sandbagging")[7] ou feindre d'être alignés avec des objectifs de sécurité tout en conservant des "préférences" cachées potentiellement néfastes ("tromperie stratégique")[8]. Détecter ces comportements est essentiel mais très complexe, et il est par conséquent très difficile de définir des seuils garantissant des niveaux de sécurité satisfaisants. Par ailleurs, l'absence d'une capacité lors d'une évaluation ne garantit pas son absence dans d'autres circonstances : ainsi, grâce à des techniques de scaffolding et de formulation optimisée des requêtes, les performances en cybersécurité d'un modèle peuvent bondir de 29% à 95% sur un même benchmark[9]. Certaines fonctionnalités n'émergeant que dans des circonstances spécifiques ou avec des méthodes de sollicitation particulières, il est particulièrement complexe de cartographier l'ensemble des capacités d'un modèle dans un cadre d'évaluation standardisé. Relever ces défis d'élicitation des capacités nécessitera la mise en place de cycles d'itération rapides entre les équipes de développement d'évaluations et les experts domaines associés à ces travaux.

[6] Voir par exemple Rauh, M., et al. Gaps in the Safety Evaluation of Generative AI. https://ojs.aaai.org/index.php/AIES/article/view/31717
[7] Voir par exemple Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, Francis Rhys Ward, "AI Sandbagging: Language Models can Strategically Underperform on Evaluations", accessible en ligne: https://arxiv.org/abs/2406.07358
[8] Voir par exemple l'analyse de l'entreprise Anthropic sur le sujet: https://www.anthropic.com/research/alignment-faking
[9] Turtayev, R., et al. Hacking CTFs with Plain Agents. https://arxiv.org/html/2412.02776v1

3. Développer une stratégie partenariale ambitieuse en Europe et à l'international

Le rôle international de l'INESIA est une question clef qu'il convient de trancher tôt dans la mise en place de l'Institut. Les homologues de l'INESIA, notamment les instituts britannique, américain, coréen ou singapourien, sont particulièrement actifs à l'international et représentent leurs gouvernements respectifs dans les conférences et débats spécialisés - un modèle dérivé d'expériences similaires en matière de diplomatie numérique et cyber.

En France, la diplomatie numérique est traditionnellement gérée directement par le ministère de l'Europe et des Affaires étrangères, qui se coordonne au niveau national avec les administrations pertinentes. Une sous-direction de la cybersécurité a par exemple été créée en 2023 au sein de la Direction Générale des Affaires politiques et de sécurité.

En ce qui concerne l'IA, il apparaît cependant nécessaire que l'INESIA puisse bénéficier d'un rôle similaire à ses homologues dans les fora internationaux. La situation du débat sur l'intelligence artificielle, dont la composante technique est essentielle y compris en matière d'accords politiques, est en effet plus comparable à des secteurs industriels complexes à enjeu dual tels que le nucléaire civil ou l'espace[10] qu'au cadre général qui a court pour la diplomatie numérique – dont l'enjeu principal consiste à adapter les normes, règles et principes existants au cyberespace et non à développer des normes d'évaluation et de contrôle, indispensables pour une technologie telle que l'IA.

La centralité de l'INESIA dans le positionnement français à l'international apparaît d'autant plus évidente que l'Institut représentera la France dans le réseau international des instituts nationaux, lancé après le Sommet de Séoul. Ce réseau doit permettre l'échange de bonnes pratiques entre homologues, renforcer la recherche sur la sécurité de l'IA, permettre le développement de missions et systèmes d'évaluation communs, de même que le développement de standards et lignes directrices à destination du secteur privé[11]. Confier à l'INESIA un mandat plus large sur de tels processus internationaux touchant à la sécurité et à l'évaluation de l'IA, dont le futur Dialogue Global sur l'IA organisé dans le cadre des Nations Unies, éviterait de multiplier les interlocuteurs pour les partenaires internationaux de la France et permettrait de centraliser l'expertise au sein d'une entité cohérente et lisible, capable de coordonner les positions avec l'ensemble des autorités publiques pertinentes et le soutien de notre réseau diplomatique.

Cette inscription internationale doit par ailleurs se développer selon une approche partenariale qui inclut l'ensemble des acteurs pertinents, y compris du secteur privé. L'INESIA doit développer un réseau de partenaires publics et privés nationaux, européens et internationaux au sein de l'écosystème de l'IA – fournisseurs, déployeurs, et évaluateurs – pour assurer un contrôle effectif et efficace des modèles. Essentiel pour développer ses capacités techniques et les mettre à l'essai sur les modèles les plus récents, cette stratégie

[10] Secteurs pour lesquels les opérateurs spécialisés participent de concert avec le ministère de l'Europe et des Affaires étrangères aux discussions internationales selon un principe de complémentarité.
[11] La lettre de mission du réseau international est disponible en ligne:
https://digital-strategy.ec.europa.eu/en/news/first-meeting-international-network-ai-safety-institutes

partenariale renforcera utilement la visibilité internationale de l'INESIA, aussi bien au sein du réseau des instituts nationaux de l'IA et dans les instances de gouvernance technique qu'auprès de l'écosystème privé.

L'INESIA doit par ailleurs veiller à ne pas dupliquer des efforts déjà entrepris par des partenaires de confiance. Collaborer étroitement avec ses homologues européens et internationaux est à ce titre essentiel, notamment en ce qui concerne le Bureau européen de l'IA. Au niveau européen, l'objectif est de prendre toute sa place dans la mise en œuvre du règlement européen sur l'IA en adoptant les instruments réglementaires nécessaires pour garantir l'accès aux modèles.

Développée en formats multilatéraux et multipartite, cette stratégie partenariale ne doit pas négliger l'importance - et la fécondité - des coopérations bilatérales. Plusieurs instituts existant ont mobilisé ces formats privilégiés pour développer des coopérations approfondies aux niveaux techniques ou de standardisation afin de faire des économies d'échelle et de renforcer leur influence commune sur des enjeux clefs de gouvernance. Ainsi par exemple des instituts britanniques et américains qui ont annoncé vouloir développer en commun les méthodes d'évaluation pour les modèles d'IA avancé[12].

[12] https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety.

Conclusion: recommandations opérationnelles pour l'INESIA

L'INESIA naît donc avec de nombreux atouts pour devenir un fer de lance en matière de promotion de l'expertise et de la vision française en matière de sécurité et d'évaluation de l'IA. Pour concrétiser cette ambition, il convient cependant de se garantir de plusieurs risques résultant tant de son mandat que des choix d'organisation qui ont présidé à sa mise en place. A ce titre, le CeSIA propose de suivre quatre axes prioritaires pour le lancement des activités de l'Institut:

1/ Prioriser le développement d'une expertise technique interne, notamment en ce qui concerne les capacités d'évaluation des modèles. Ce travail doit notamment s'appuyer sur l'écosystème français, européen et international de confiance existant.

2/ Structurer l'INESIA dans une logique de spécialisation afin d'éviter les luttes de service, en sanctuarisant notamment son personnel et un budget dédié pour les activités de recherche et d'évaluation.

3/ Développer une politique de partenariats en France et en Europe avec les acteurs privés en soutien au développement d'une méthodologie d'évaluation de référence

4/ Positionner l'INESIA comme le point de contact central en France pour les discussions internationales ayant trait à la sécurité et à l'évaluation de l'IA.


Centre pour
la Sécurité de l'IA
www.securite-ia.fr