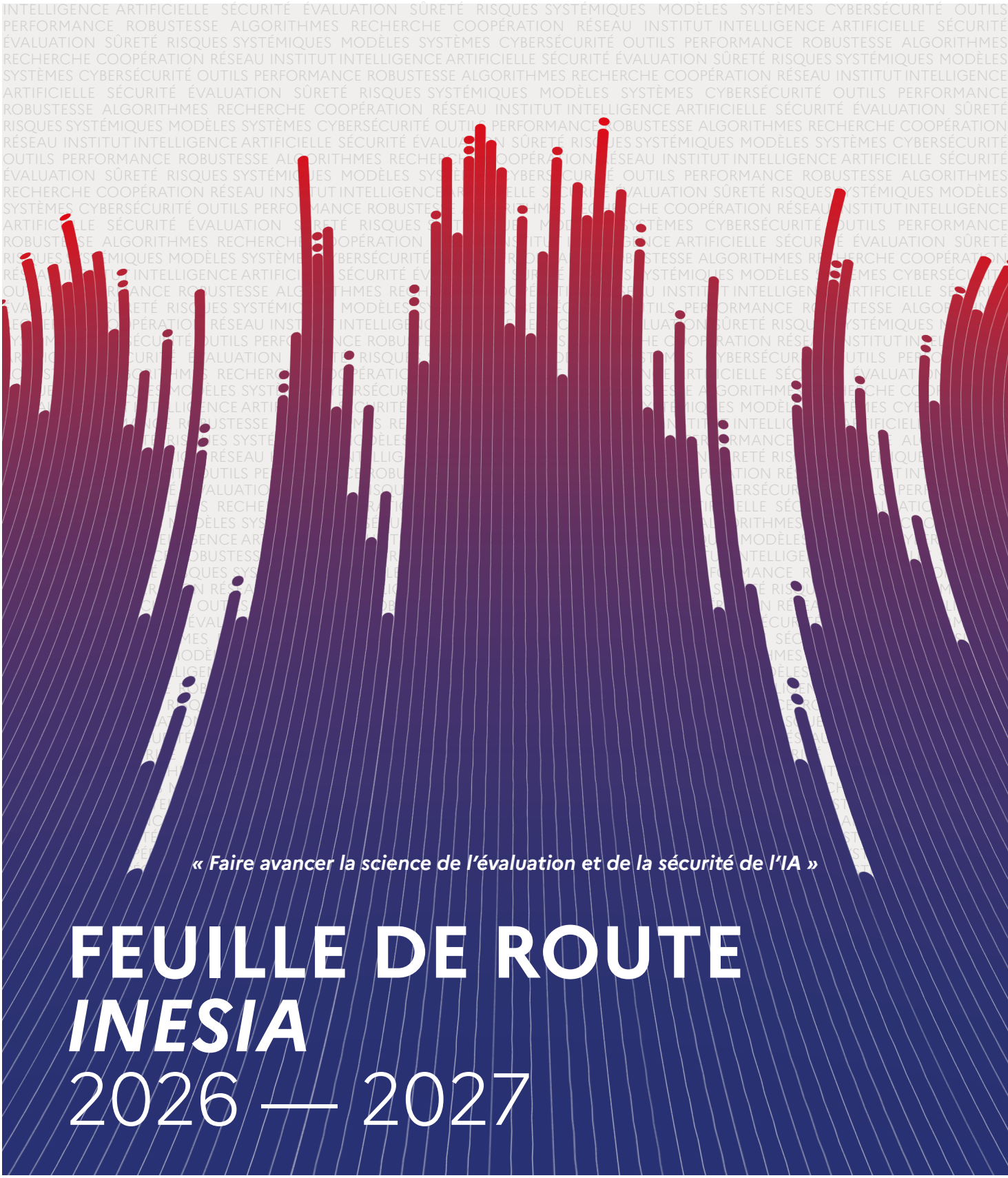
GOUVERNEMENT

Liberté
Égalité
Fraternité



« Faire avancer la science de l'évaluation et de la sécurité de l'IA »

FEUILLE DE ROUTE *INESIA* 2026 — 2027

RÉSUMÉ ÉXÉCUTIF

Les systèmes d'IA franchissent de nouveaux seuils de complexité. Dotés de capacités de raisonnement multimodal, capables d'interagir avec des environnements numériques et physiques, voire de s'adapter en continu, les modèles d'IA les plus avancés s'imposent dans tous les champs de la société. Déjà intégrés dans les services publics, la recherche, l'éducation ou les outils professionnels, ils sont appelés à participer à des décisions aux conséquences sociales, économiques et sécuritaires majeures.

Ces évolutions font émerger une série de défis inédits. La capacité des IA à générer des contenus indiscernables du réel, à influencer des comportements humains ou à orchestrer des actions dans des systèmes complexes accentue leur potentiel systémique. Conçues pour des systèmes fermés, statiques et monofonctionnels, les méthodes d'évaluation traditionnelles apparaissent face à cela dépassées.

L'évaluation de l'IA devient donc un enjeu stratégique à part entière. Elle ne se limite pas à mesurer une performance : elle doit permettre de détecter des comportements inattendus, de quantifier des incertitudes, d'évaluer la robustesse, l'alignement, la résilience et les risques d'usage détourné. Elle mobilise un champ interdisciplinaire exigeant, combinant science des données, modélisation des comportements, sécurité informatique, ergonomie cognitive et philosophie des techniques.

Dans ce paysage mouvant, à l'intersection des dynamiques technologiques et des exigences de souveraineté, a été créé l'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA). Lancé en 2025, il regroupe l'ANSSI, Inria, le LNE et le PEReN. Il a pour mission de structurer une capacité publique d'évaluation des IA avancées, à la fois rigoureuse, indépendante et au besoin interopérable avec les initiatives européennes et internationales.

L'INESIA agit à la croisée de trois dimensions : le soutien à la régulation en appui à la mise en œuvre du règlement européen sur l'intelligence artificielle (RIA), la maîtrise des risques systémiques de l'IA et le soutien à l'évaluation de la performance et de la fiabilité des modèles et systèmes. Son action est conduite avec méthode : comprendre les systèmes émergents, tester leurs comportements, structurer les outils d'évaluation, influencer les normes et orienter les choix collectifs.

En poursuivant la présente feuille de route, l'INESIA entend poser les bases d'une capacité d'évaluation de l'IA à la fois robuste, ouverte et souveraine, en mesure d'éclairer les décisions publiques, de renforcer la résilience collective face aux risques et de garantir que l'innovation technologique s'inscrive dans un cadre de confiance.

SOMMAIRE

INTRODUCTION

CONSTAT

Le rythme des progrès de l'IA ne faiblit pas. Le développement de modèles capables de raisonnement, de manipulation de données multimodales, d'interactions avec des environnements hybrides, et prochainement d'auto-amélioration pose des défis nouveaux. Ils interrogent sur les plans de la performance et de l'interprétabilité, comme de la sûreté, de la robustesse face à l'imprévu, ou de l'alignement aux valeurs et intentions humaines.

Les méthodes d'évaluation conçues pour des systèmes fermés, monofonctionnels ou statiques, montrent leurs limites. Il est nécessaire que soient développés de nouveaux référentiels pour évaluer leur robustesse, leur résilience, leur fiabilité contextuelle ou encore les risques que poseraient des usages détournés. Alors que **l'évaluation de l'IA devient un champ scientifique à part entière**, il est fondamental que la France dispose de compétences techniques pour être capable d'identifier, mesurer et contenir les risques à fort potentiel systémique, y compris ceux liés à la manipulation de l'information, à la cybersécurité ou à la déstabilisation de processus collectifs.

RÉPONSE

L'INESIA se donne pour ambition de contribuer à cette « science de l'évaluation ». Créé en 2025, l'INESIA a pour mission de bâtir une capacité souveraine et publique d'évaluation des IA avancées, en structurant un socle de savoirs rigoureux, en fédérant les acteurs concernés, et en facilitant les échanges entre monde académique, administration, industrie et société civile. Il vise à structurer une capacité souveraine d'évaluation de l'IA, à l'intersection de l'excellence scientifique, de l'innovation technique et de l'exigence vis-à-vis de son impact opérationnel.

L'INESIA assure la mise à disposition d'une expertise technique aux services de l'État et favorise une innovation responsable. Il constitue un point d'ancrage national dans les réseaux internationaux de confiance, et à ce titre assure la représentation de la France au sein du réseau des AI Safety Institutes.

L'action de l'INESIA repose sur quatre principes fondamentaux : **construire une démarche scientifique et stratégique, structurer une capacité souveraine d'évaluation, coordonner les expertises publiques et mutualiser les moyens, être lisible et actif dans les réseaux internationaux**. Sur le plan opérationnel, cette démarche implique de mieux comprendre les systèmes à évaluer (via de nouvelles métriques, de nouveaux tests), tester plus loin et plus vite (via des outils et plateformes techniques solides), diffuser ces connaissances à la fois dans la communauté scientifique et dans l'écosystème normatif et orienter les pratiques d'évaluation des futurs modèles les plus puissants. Cette ambition s'organise autour de trois pôles stratégiques et un axe transverse.

PÔLE 1 – APPUI À LA RÉGULATION

POUR CONTRIBUER A UNE MISE EN ŒUVRE EQUILIBREE ET EFFICACE DU CADRE REGLEMENTAIRE DE L'IA, QUI PROTEGE LES CITOYENS ET SOUTIENNE L'INNOVATION.

L'INESIA veillera à contribuer à la mise à disposition des autorités d'outils d'évaluation nécessaires à leurs activités (**Projet n°1**.). Des points de situation réguliers permettront de suivre les travaux normatifs accompagnant la mise en œuvre du RIA (**Projet n°2**.). Par ailleurs, les travaux déjà lancés dans le champ de la détection de contenus synthétiques seront poursuivis (**Projet n°3**.). Enfin, des méthodes d'évaluation adaptées à la cybersécurité des systèmes d'IA et des produits de cybersécurité intégrant de l'IA seront développées (**Projet n°4**.).

PÔLE 2 – RISQUES SYSTÉMIQUES

POUR OBJECTIVER LES RISQUES SYSTEMIQUES QUE POURRAIT POSER L'IA, INFORMER LA DECISION PUBLIQUE ET AFFIRMER LA PLACE DE LA FRANCE DANS LA SECURITE DE L'IA.

Des activités de recherches seront conduites afin de mieux comprendre les risques systémiques, les caractériser, les maîtriser et informer la décision publique (**Projet n°5**.). Plus en aval, des travaux permettront d'affiner l'évaluation des systèmes agentiques (**Projet n°6**.). Enfin, la poursuite des activités conduites dans le cadre du réseau international des AISI affirmera la place de la France dans les efforts en faveur de la sécurité de l'IA (**Projet n°7**.).

PÔLE 3 – PERFORMANCE ET FIABILITÉ

POUR STIMULER LA CREATIVITE DE L'ECOSYSTEME, FAIRE DE L'EMULATION UN LEVIER POUR FAVORISER L'EMERGENCE DE SOLUTIONS PLUS PERFORMANTES ET PLUS FIABLES.

Des challenges seront conduits, qui permettront de renforcer l'émulation entre les acteurs et favoriseront le développement de solutions créatives (**Projet n°8**.).

AXE TRANSVERSE

POUR CREER DES COMMUNS STRUCTURANTS POUR LES TRAVAUX DE L'INESIA, DEVELOPPER LES SYNERGIES DES PARTIES ET DE LEURS TRAVAUX, FEDERER UN ECOSYSTEME.

Une mise en commun des bonnes pratiques et connaissances permettra de capitaliser sur l'existant et de renforcer les connaissances (**Projet n°9**.). Il sera nécessaire de mettre en place l'architecture technique qui permettra l'évaluation des modèles et systèmes (**Projet n°10**.). Finalement, l'organisation d'activités d'animation scientifique permettra de fédérer l'écosystème de l'évaluation et de la sécurité de l'IA (**Projet n°11**.).

Selon cette vision, la présente feuille de route décline les priorités de l'INESIA en actions concrètes et échelonnées dans le temps. Elle a été élaborée en articulation avec les besoins identifiés par les administrations concernées, les orientations européennes, les initiatives internationales en cours et les travaux scientifiques de référence.

PÔLE APPUI À LA RÉGULATION DE L'IA

Contribuer à une mise en œuvre efficace du cadre réglementaire de l'IA, pour une application équilibrée qui protège les citoyens et soutienne l'innovation.

L'INESIA a vocation à mettre une expertise technique à disposition des autorités de régulation de l'IA. Face à la complexité croissante des systèmes, à la diversité de leurs usages et aux exigences réglementaires, il est essentiel de concevoir des protocoles de test systématiques, réplicables et adaptés aux besoins concrets des autorités publiques.

Le Projet n°1. veille à contribuer à la **mise à disposition d'outils d'évaluation à destination des autorités de surveillance du marché** qui assureront le contrôle des systèmes d'IA à haut risque dans le cadre de la mise en œuvre du RIA. Dans une perspective d'efficacité, les outils existants seront listés et rassemblés, et des travaux de recherche pourront être entrepris afin de développer les outils manquants.

Une **acculturation des membres de l'INESIA aux questions normatives**, notamment dans le champ de l'évaluation, permettra d'échanger autour des réflexions en cours au sein des différentes instances de normalisation : ISO-IEC, CEN-CENELEC, etc. (Projet n°2.).

Le Projet n°3. poursuivra les travaux déjà lancés qui évaluent avec précision la performance de **détecteurs de contenus synthétiques dans des conditions réelles**, enjeu clé pour garantir la sécurité de l'information. Dans ce prolongement seront renforcées les capacités de détection de contenus générés par IA, en élargissant et fiabilisant une bibliothèque open-source de référence.

Enfin, le projet SEPIA (SEcurité des Produits intégrant de l'IA, Projet n°4.) permettra de **développer des méthodes de certification adaptées à la cybersécurité des systèmes d'IA et des produits de cybersécurité intégrant de l'IA**, à destination de l'écosystème de certification, et de façon à proposer une réponse harmonisée aux différents cadres réglementaires concernés (règlement européen sur l'IA, règlement européen sur la cyber-résilience, règlement européen sur la cybersécurité).

PROJET N°1.

METTRE DES OUTILS D'ÉVALUATION À LA DISPOSITION DES RÉGULATEURS NATIONAUX

OBJECTIF

Mettre des outils d'évaluation à disposition des autorités françaises de surveillance du marché au titre du règlement IA, par la cartographie des outils existants pertinents d'une part, par le développement d'outils manquants en lien avec leurs besoins de l'autre.

CONTEXTE

Dans le cadre de la mise en œuvre du règlement IA, plusieurs autorités compétentes (dites de surveillance du marché) seront chargées de contrôler les systèmes d'IA à haut risque mis en service sur le marché français. Elles vérifieront la conformité de ces produits selon différentes exigences techniques et administratives.
Les autorités compétentes en France bénéficieraient d'un recensement des outils d'évaluation, dans un souci d'efficience et d'harmonisation des pratiques, et d'un cadre de mise au point des méthodes manquantes.

MÉTHODOLOGIE

En sa qualité de membre du socle de compétences techniques mutualisé, le PEReN prévoit d'effectuer un recueil de besoins auprès des autorités de surveillance de marché. Une prise de contact avec les homologues européens pourra aussi être proposée. Les enseignements tirés de ces échanges seront versés au projet. En parallèle, un recensement des outils et moyens existants sera effectué par les membres de ce projet et aboutira, selon les lacunes identifiées compte tenu des besoins, à l'identification d'axes de recherche et développement prioritaires. Les outils et méthodologies d'évaluation prioritaires manquants pourront faire l'objet d'activités de recherche ou de développement, en lien avec les normes harmonisées prévues en soutien au RIA.
Le projet pourra être conduit en lien avec le Projet n°10. en tant que de besoin.

LIVRABLES

- Synthèse des besoins des régulateurs et identification des points nécessitant le concours de l'INESIA à travers des travaux de recherche et développement.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN.

PROJET N°2.

FAVORISER LES ÉCHANGES SUR LES QUESTIONS DE NORMALISATION DE L'IA

OBJECTIF

Favoriser les échanges entre les membres de l'INESIA sur les sujets de normalisation, notamment en lien avec l'évaluation et la sécurité des modèles d'IA à usage général, ainsi qu'à la réponse à la demande de normalisation en soutien au RIA.

CONTEXTE

La Commission européenne a ouvert une demande de normalisation en soutien aux exigences de conformité du règlement IA relatives aux systèmes d'IA à haut risque, dont les travaux sont toujours en cours. Par ailleurs, elle soumettra également une demande de normalisation relative aux modèles d'IA à usage général, incluant des questions telles que leur sécurité, leurs modalités d'évaluation, et leur empreinte environnementale.
D'autres cadres tels que l'OCDE, le réseau des AISI et le processus d'Hiroshima travaillent à définir des cadres communs sur les tests de robustesse, de fiabilité, de transparence ou de cybersécurité des systèmes d'IA.

MÉTHODOLOGIE

Fortement impliquée dans le suivi de la mise en œuvre du règlement IA et mobilisée sur le suivi des travaux normatifs, la DGE pourra organiser des points de situation consacrés aux normes harmonisées afin d'assurer un partage de l'information entre les membres de l'INESIA. Elle pourra par ailleurs alerter en tant que de besoin les différentes parties prenantes de points normatifs d'intérêt, en dehors de ce cycle de réunions.
Selon les opportunités, ces travaux pourront inclure de manière plus large les autres parties prenantes publiques, actives au sein des enceintes de normalisation.

LIVRABLE

- Organisation de réunions trimestrielles visant à vulgariser le travail en cours et à informer les parties prenantes des avancées normatives.

PARTIES PRENANTES AU PROJET

DGE, ANSSI, LNE et PEReN.

PROJET N°3.

ÉVALUER LA DÉTECTION DE CONTENUS ARTIFICIELS EN CONDITIONS RÉELLES

OBJECTIF

Renforcer les capacités d'évaluation des détecteurs de contenus générés artificiellement (texte, image, audio, vidéo) par le développement d'une bibliothèque de référence, régulièrement actualisée, permettant de comparer les performances des outils de détection dans des conditions proches du terrain.

CONTEXTE

Les avancées de l'IA générative se traduisent par une diffusion rapide de contenus synthétiques bientôt indiscernables à l'œil de contenus authentiques, qui posent un risque sur la confiance dans l'information : désinformation, ingérence, manipulation de masse. Il est stratégique pour les autorités publiques de disposer d'une capacité à les détecter.

La majorité des travaux existants reposent pourtant sur des tests en laboratoire peu représentatifs de la diversité des contenus réellement diffusés sur Internet. En 2024, le PEReN et VIGINUM ont amorcé le développement d'une bibliothèque open-source dédiée à l'évaluation des détecteurs de textes et images générés.

Diffusé en source ouverte dans le cadre du Sommet de Paris sur l'IA, ce socle constitue une première avancée. Il s'agit désormais de le maintenir en assurant une veille active pour garantir sa pertinence scientifique et opérationnelle, et de l'étendre à d'autres modalités critiques comme l'audio ou la vidéo.

MÉTHODOLOGIE

Le projet comprendra trois volets principaux :

- veille scientifique et revue d'état de l'art : recensement et analyse critique des méthodes de détection de contenus audio et vidéo générés artificiellement ;
- extension de la bibliothèque de détection : intégration de nouvelles modalités (audio, vidéo) et ajout de détecteurs pertinents issus de l'état de l'art, dans le prolongement du code existant codéveloppé par le PEReN et VIGINUM ;
- si opportun, exploration d'approches ensemblistes combinant plusieurs outils dans un souci de performance.

Au besoin, des jeux d'évaluation non publics pourront être constitués.

LIVRABLES

- Etat de l'art actualisé sur la détection de contenus générés ;
- Base de code enrichie ;
- Résultats de l'évaluation sur jeux de données non publics à des fins de benchmarks dédiés, i.e. en conditions réelles.

PARTIES PRENANTES AU PROJET

ANSSI et PEReN. En partie extérieure, VIGINUM.

PROJET N°4.

MÉTHODES D'ÉVALUATION DE LA CYBERSÉCURITÉ DES SYSTÈMES D'IA ET DES PRODUITS DE CYBERSÉCURITÉ INTÉGRANT DE L'IA (SEPIA)

OBJECTIF

Élaborer des méthodes d'évaluation adaptées à la cybersécurité des systèmes d'IA et des produits de cybersécurité intégrant de l'IA, à destination de l'écosystème afin de renforcer la sécurité des systèmes d'intelligence artificielle, et de façon à proposer une réponse harmonisée aux différents cadres réglementaires concernés (RIA, CRA, CSA). L'objectif de ces travaux est également de porter et valoriser ces travaux dans les instances européennes et internationales traitant de la certification de cybersécurité.

CONTEXTE

En plus des vulnérabilités de cybersécurité usuelles, les systèmes d'IA sont exposés à des vulnérabilités spécifiques liées aux modèles d'IA. L'impact de ces vulnérabilités et les moyens d'y remédier doivent être considérés non seulement à l'échelle des composants individuels, mais également du point de vue de l'architecture du système d'IA et de son intégration au sein d'un système d'information. Pourtant, il n'existe pas aujourd'hui de cadre d'évaluation et de certification permettant d'évaluer la cybersécurité des systèmes d'IA. Si le RIA prévoit des évaluations de conformité des systèmes d'IA notamment sur les aspects cyber, il convient de construire des méthodes d'évaluation permettant d'attester du niveau de cybersécurité d'un système d'IA à travers des évaluations et des tests de pénétration réalisés par un évaluateur tiers.
Pour y répondre, l'ANSSI pilote depuis début 2025 le projet SEPIA (SEcurité des Produits IA), qui vise à anticiper ces mutations à travers l'élaboration de méthodes spécifiques d'évaluation, en lien avec l'écosystème français et les partenaires étrangers.

MÉTHODOLOGIE

Le projet s'appuie sur un groupe de travail inter-agences (ANSSI, Inria, LNE, AMIAD, Centre d'évaluation de la sécurité des technologies de l'information CESTI) qui se réunit mensuellement afin de développer les livrables souhaités.
Selon l'opportunité, les travaux pourraient inclure des cas pratiques d'évaluation des méthodes d'évaluation développées, au travers d'expérimentation qui impliqueraient les CESTI, les parties de l'INESIA et des acteurs industriels. Un même produit pourrait être testé par plusieurs entités ou bien plusieurs produits pourraient être évalués simultanément. Selon l'opportunité également, des actions de sensibilisation auprès des autorités de surveillance des marchés en application du RIA pourraient être conduites.
Le projet pourra être conduit en lien avec le Projet n°5. en tant que de besoin.

LIVRABLES

- Méthodes d'évaluation pour la cybersécurité des systèmes et produits embarquant de l'IA ;
- Recensement des attaques connues (software, hardware), et identification de l'effort nécessaire à leur mise en œuvre (niveau d'analyse de vulnérabilité).

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE. CESTI (Almond, CEA-Leti, Lexfo, Oppida, Quarkslab, Thales/CNES, Serma Safety & Security), AMIAD.

PÔLE RISQUES SYSTÉMIQUES

Objectiver et évaluer les risques systémiques que pourrait poser l'IA, informer la décision publique et affirmer la place de la France dans la sécurité de l'IA.

Les progrès technologiques et le déploiement de l'IA sont le moteur de changements rapides de la société et soulèvent des préoccupations quant aux risques systémiques qui pourraient émerger. Incarnant l'engagement français au sein du réseau des AI Safety Institutes, l'INESIA travaillera à éclairer la décision publique et contribuer à un cadre de confiance propice à la diffusion de l'IA.

Par la définition de seuils mesurables et la modélisation spécifique des scénarios critiques, le Projet n°5. vise à **objectiver les risques systémiques** en lien avec les problématiques de manipulation d'information à grande échelle (santé, système financier, systèmes d'information), de cybersécurité offensive et de prolifération NRBC. Il proposera également des méthodes d'atténuation.

Plus en aval, le Projet n°6. se concentre sur **l'évaluation des systèmes d'agents**, en développant des protocoles destinés à tester leurs capacités de raisonnement, d'interaction et de convergence dans des scénarios critiques.

Le Projet n°7. affirme la place de la France dans **le réseau des AI Safety Institutes.** Par une participation à ses travaux et notamment aux évaluations conjointes, elle confirmera sa place dans les efforts mondiaux pour la sécurité de l'IA et renforcera sa propre expertise. Si dans la vague de travaux actuelle, les activités conduites s'incarnent essentiellement dans le Projet n°6, le périmètre sera redéfini à l'occasion selon le programme des prochaines vagues.

PROJET N°5.

DÉVELOPPER UNE EXPERTISE TECHNIQUE SUR L'ÉVALUATION ET L'ATTÉNUATION DES RISQUES SYSTÉMIQUES : MANIPULATION DE L'INFORMATION, CYBERSÉCURITÉ OFFENSIVE, NRBC

OBJECTIF

Le projet s'inscrit dans une démarche de recherche sur l'évaluation de l'impact de l'IA dans le domaine des risques systémiques. L'enjeu central est ici de comprendre comment les modèles d'IA disposant de capacités à fort impact peuvent entraîner des effets négatifs réels sur la santé publique, la sûreté, la sécurité publique, les droits fondamentaux ou la société dans son ensemble, pouvant être propagés à grande échelle tout au long de la chaîne de valeur.
Le projet s'appuiera sur une analyse des seuils technologiques et comportementaux critiques, susceptibles d'amplifier ou de déclencher des effets systémiques. Il visera à concevoir des protocoles d'évaluation et de détection précoce exploitables tant par les développeurs (pour renforcer la sécurité des modèles) que par les acteurs de la régulation (pour anticiper et encadrer les usages à risque). Enfin, le projet ambitionne de fournir aux décideurs publics une capacité d'anticipation et un éclairage prospectif, fondés sur des indicateurs de vulnérabilité et des scénarios d'impact. Cet appui participera à l'élaboration de politiques de prévention, de supervision et de gouvernance adaptées à l'évolution rapide des capacités et des usages de l'IA.

CONTEXTE

Certains usages de l'IA pourraient engendrer des effets de grande ampleur difficilement prévisibles, exploitables à des fins malveillantes : désinformation et attaques contre la cybersécurité, prolifération, déstabilisation d'un système économique, etc. Les progrès technologiques abaissent les barrières à l'entrée du développement de modèles avec un fort potentiel de nuisance, délibéré ou accidentel.
Face à ces menaces, les cadres actuels d'évaluation restent insuffisants. Il devient nécessaire de mieux comprendre comment évaluer les risques liés à l'utilisation de l'IA, d'identifier les seuils de capacité au-delà desquels un système peut devenir dangereux, et de formaliser des recommandations pour les auditer et les contenir.

MÉTHODOLOGIE

Le projet relève principalement d'activités de recherche fondamentale et appliquée. Articulées autour de trois axes thématiques complémentaires, elles visent à développer la méthodologie d'évaluation de l'IA pour caractériser, tester et anticiper des formes distinctes de risques systémiques associés aux modèles d'IA avancés, tout en développant des approches d'atténuation :

- Évaluer la capacité des modèles d'IA à contribuer (intentionnellement ou non) à des risques, notamment dans le champ NRBC. Sur ce dernier champ, conduite d'uplift studies, identification de seuils critiques, conception de bancs d'essai pour détecter le franchissement de ces seuils ;
- Analyser la capacité de persuasion et de manipulation d'agents conversationnels sur des utilisateurs humains, pour mieux comprendre les conditions d'émergence d'une manipulation de masse et concevoir des outils de vigilance et d'alerte. Par exemple, étudier la capacité de conviction d'agents conversationnels sur les humains, détecter une tentative de manipulation de masse par une identification des contenus générés, des manières de le générer, de la campagne elle-même ;
- Analyser les capacités de réalisation d'attaques informatiques, entendu que l'IA permet de générer des actions ciblant la confidentialité, l'intégrité ou la disponibilité des systèmes d'information. Le niveau de menace exact correspondant n'ayant pas été évalué rigoureusement à ce jour, des travaux de recherche restent nécessaires pour répondre à diverses questions d'intérêt.

En parallèle de ces travaux sur l'évaluation, le projet développera des méthodes d'atténuation. Il proposera notamment des garde-fous intégrés au niveau des systèmes d'IA eux-mêmes (filtrage des données d'entraînement ou de fine-tuning, recommandations d'entraînement et d'optimisation, détection d'abus, vérification de cohérence interne), et recherchera des contre-mesures défensives renforçant la résilience des infrastructures et applications ciblées par les usages malveillants de l'IA.

Le projet pourra être conduit en lien avec d'autres projets en tant que de besoin, notamment avec le Projet n°4.

LIVRABLES

- Identification des projets de recherche spécifiquement ciblés sur l'évaluation et l'atténuation des risques systémiques ;
- Publication de méthodes d'atténuation associées à des outils en source ouverte si opportun ;
- Constitution de jeux d'évaluation non publics si opportun.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN. En partie extérieure, VIGINUM et AIST (SGDSN).

PROJET N°6.

ÉVALUER LES PERFORMANCES ET LES RISQUES DES SYSTÈMES AGENTIQUES

OBJECTIF

Développer une méthodologie d'évaluation rigoureuse des systèmes d'agents IA fondés sur des modèles de langage. Caractériser notamment leurs capacités en matière de cybersécurité et estimer la mesure des possibilités d'usage à des fins criminelles en vue notamment d'informer les pouvoirs publics et d'affiner l'anticipation des risques. Contribuer aux travaux conduits dans le cadre du réseau des AI Safety Institutes, de façon à affirmer la place de la France dans le réseau.

CONTEXTE

Les architectures d'agents constituent aujourd'hui l'un des principaux relais de renforcement des capacités de l'IA et un vecteur majeur de son adoption à large échelle. En 2024, l'écosystème a vu émerger une nouvelle génération de systèmes agentiques, combinant LLMs, outils externes, mémoire de long terme et capacités de raisonnement ou de planification. Capables de décomposer une tâche, mobiliser des outils et interagir avec un environnement complexe, ils posent des défis inédits :

- le résultat final ne reflète qu'en partie la qualité du raisonnement sous-jacent ;
- les comportements dépendent fortement des conditions d'expérimentation ;
- certains usages critiques pour la sécurité doivent être anticipés via des scénarios ciblés.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN.

MÉTHODOLOGIE

Le projet commencera par un état de l'art sur les risques identifiés et les critères et protocoles d'évaluation orientés sécurité. Cet état de l'art pourra être enrichi par des contributions dans le cadre de ce projet. La liste de risques identifiés pourra aussi être nourrie par des cas d'usages supplémentaires identifiés par les partenaires du projet. L'objectif est de lister d'une part les typologies de risques potentiels à travers des scénarios critiques, et d'autre part les critères et protocoles d'évaluation retenus pour l'INESIA. Dans un deuxième temps ou par itération successives, il s'agira d'implémenter ces protocoles, en s'appuyant, quand cela est pertinent, sur l'existant. Cela fera notamment l'objet d'une revue des bibliothèques d'évaluation existantes, de développements nouveaux si nécessaire, ainsi que l'utilisation ou la création de jeux de données de référence dédiés.

LIVRABLES

- Rapport sur les risques et les protocoles d'évaluation orientés sécurité dans l'état de l'art, permettant de sélectionner certains scénarios critiques ;
- Identification ou développement d'une bibliothèque d'évaluation qui implémente ces protocoles ;
- Spécifications puis identification ou création de jeux de données d'évaluation.

PROJET N°7.

PRENDRE PART AUX TRAVAUX DU RÉSEAU DES AI SAFETY INSTITUTES

OBJECTIF

Contribuer aux travaux conduits dans le cadre du réseau des AI Safety Institutes, de façon à renforcer l'expertise de l'INESIA dans le champ de l'évaluation et à affirmer la place de la France dans le réseau.

CONTEXTE

Le réseau des AI Safety Institutes permet une coopération internationale dans le champ de l'évaluation de l'IA. À l'occasion du Sommet pour l'action sur l'IA organisé par la France les 10 et 11 février 2025, les représentants de plusieurs parties du réseau international des AI Safety Institutes (AISI) réunis à Paris ont dévoilé les premiers résultats d'essais menés conjointement. La France y a notamment fourni le jeu de données ayant permis l'évaluation portant sur la robustesse cyber des modèles et a réalisé les évaluations de performance relatives aux risques de sécurité lorsque l'on interagit avec le modèle en langue française.
Par ailleurs, les travaux conduits à l'occasion des trois premières vagues d'évaluation donneront lieu à une valorisation à l'occasion de la conférence ICML à Vancouver. En lien avec le projet correspondant de la présente feuille de route, ils concernaient l'évaluation des systèmes agentiques.

MÉTHODOLOGIE

Par nature, le contenu précis de ce projet restera à définir en fonction des orientations fixées au sein du réseau des AI Safety Institutes. Les parties de l'INESIA pourront proposer certaines thématiques en lien avec leurs activités, que le SGDSN pourra porter auprès du réseau à l'occasion de la définition du programme des futures vagues de travail.

LIVRABLES

- Travaux d'évaluation conjoints ou autres, selon le programme de travail du réseau ;
- Si pertinent, publications scientifiques et communications.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN.

PÔLE PERFORMANCE ET FIABILITÉ

Stimuler la créativité de l'écosystème, faire de l'émulation un levier pour favoriser l'émergence de solutions plus performantes, plus fiables.

Ce troisième pôle de l'INESIA contribuera aux progrès des niveaux de fiabilité et de performance des modèles et systèmes d'IA, dans la mesure où des considérations d'intérêt général justifient une implication de la puissance publique.

L'organisation de challenges constitue un levier majeur afin de stimuler l'écosystème autour de thématiques d'intérêt pour l'évaluation ou la sécurité de l'IA. En parvenant à susciter un esprit de « coopétition », l'INESIA parviendrait à faire émerger des solutions créatives. Le Projet n°8. invite à proposer des pistes et des cadrages pour de tels challenges.

PROJET N°8.

ORGANISATION DE CHALLENGES

OBJECTIF

Les challenges ont pour ambition de susciter une émulation internationale permettant une montée en maturité rapide de technologies d'IA, au travers d'une campagne d'évaluation ouverte à l'international. Leur enjeu consiste à créer un effet d'entraînement important pour clarifier et faire progresser l'état de l'art. Il s'agit de susciter une « coopétition » entre les consortia participants, créant une émulation fédératrice et structurante qui favorise les échanges entre experts, assurant leur mobilisation, maximisant l'innovation et les progrès technologiques réalisés.

CONTEXTE

L'évaluation de l'IA fait face à de nombreux enjeux techniques et méthodologiques pour permettre une adoption large et sereine dans les entreprises et dans la société : fiabilité, gestion de la multimodalité, personnalisation du contenu, adaptation à un contexte ou à une terminologie spécifique, etc. L'organisation de challenges permet d'opérer des vérifications de la performance et de la fiabilité des IA, dans une dynamique à la fois collaborative et comparative, permettant à la fois de faire progresser la science de l'évaluation sur les cas concrets apportés par les participants aux challenges, tout en stimulant l'innovation.

MÉTHODOLOGIE

Les challenges seront structurés afin de susciter une émulation internationale et permettre à différents acteurs de l'écosystème IA de se mobiliser, pour résoudre une tâche d'évaluation spécifique de l'intelligence artificielle. Afin de pouvoir attirer des participants diversifiés, aussi bien français qu'internationaux, les challenges devront se concentrer sur des problématiques opérationnelles, être facilement déployables et permettre d'aboutir rapidement à des résultats concrets. Un volet consacré à l'animation scientifique (Projet n°11.) pourra également constituer un vecteur pour la communication et la dissémination des travaux de l'INESIA. Les modalités du challenge retenu seront précisées en concertation avec le comité stratégique de l'INESIA, lors de l'élaboration du cahier des charges.

LIVRABLES

Les premiers travaux de ce projet se concentreront sur l'organisation d'un challenge pilote, qui servira de base aux challenges INESIA ultérieurs.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN.

AXE TRANSVERSE

Créer des communs structurants pour les travaux de l'INESIA, développer les synergies des parties et de leurs travaux, fédérer un écosystème.

Au-delà des activités conduites dans le cadre des trois pôles, un ensemble de travaux sont nécessaires au bon fonctionnement et au rayonnement de l'INESIA. En se dotant d'un cadre clair et d'outils communs, en partageant savoirs et bonnes pratiques, les parties renforceront leurs synergies et accéléreront leur montée en expertise.

L'enjeu premier est de capitaliser sur les connaissances actuelles, et de les compléter en un véritable corpus de connaissances sur l'évaluation de l'IA. Le Projet n°9. visera à recenser les méthodes, outils et concepts disponibles, à identifier les lacunes, et à mettre à disposition des autorités politiques une synthèse claire et dynamique des savoirs mobilisables.

Il est par ailleurs nécessaire que l'INESIA dispose de moyens d'évaluation qui répondent à ses besoins, capable notamment de supporter l'évaluation des modèles sensibles et de participer aux exercices coordonnés menés par les AI Safety Institutes (Projet n°10.).

Enfin, un ensemble d'activités de communication et de veille scientifique permettront de mettre en valeur les travaux de l'INESIA et d'approfondir les échanges sur l'évaluation et la sécurité de l'IA au niveau national, voire international. En facilitant les échanges dans une mesure large englobant écosystèmes privés comme public, en diffusant les bonnes pratiques et en mettant en valeur les travaux de l'INESIA, il participera à structurer la communauté (Projet n°11.).

PROJET N°9.

POSER UN CADRE DE MISE EN COHÉRENCE DES ACTIVITÉS DE VEILLE ACADÉMIQUE ET MÉTHODOLOGIQUE

OBJECTIF

Plusieurs des travaux de l'INESIA reposent sur des activités de veille : les Projet n°1. (Mettre des outils d'évaluation à la disposition des régulateurs nationaux) et Projet n°10. (Assurer l'accès à une solution d'évaluation de l'IA) requièrent ainsi un suivi des avancées méthodologiques de l'évaluation, quand le Projet n°11. (Animation scientifique autour des travaux de l'INESIA) implique d'assurer un suivi des publications académiques. Le projet n°9. visera à centraliser, uniformiser, articuler les connaissances acquises par les partenaires INESIA durant l'exécution de leurs travaux.
Ces activités fourniront une ressource centrale pour l'évaluation IA et offriront un appui aux décisions stratégiques concernant les efforts de recherche et développement dédiés à l'évaluation de l'IA.

MÉTHODOLOGIE

Le projet proposera un cadre et une base technique pour centraliser et mettre en relation les résultats des travaux de revue et de veille, sur les plans académique comme méthodologique. Ces états de l'art et veilles pourront concerner par exemple les connaissances acquises relatives aux benchmarks, métriques, méthodes d'évaluation et outils. Ces travaux s'appuieront sur l'analyse de la littérature ainsi que des résultats produits par la communauté scientifique et industrielle.

CONTEXTE

Alors qu'il n'existe pas à date de consensus sur les méthodes d'évaluation des modèles et systèmes d'IA avancés, les travaux du réseau des AISI ont mis en évidence des écarts significatifs entre les évaluations réalisées par différents pays, pour des modèles évalués pourtant dans des conditions proches.
La conduite des activités de l'INESIA suppose un travail à l'état de l'art, dans des domaines variés. À date, les parties ont déjà établi des bases de connaissance, qui ne peuvent être exploitées par les autres, faute d'avoir déjà été partagées. De façon à maximiser leur utilité et à mettre à disposition de l'INESIA un vrai référentiel, il est nécessaire de poser un cadre qui mette en cohérence ces connaissances.
En tant que de besoin, des articulations avec les autres projets, notamment Projet n°1., Projet n°10., Projet n°11. seront considérées.

LIVRABLE

- Mutualisation des travaux de veille au sein de l'INESIA.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN.

PROJET N°10.

ASSURER L'ACCÈS À UNE SOLUTION D'ÉVALUATION DE L'IA

OBJECTIF

Concevoir, développer et déployer une infrastructure logicielle et matérielle adaptée à l'évaluation des systèmes d'IA à partir des cas d'usage fournis par l'INESIA. Cette plateforme modulaire sera capable de mener des évaluations reproductibles dans le cadre de protocoles interopérables, d'assurer des tests sécurisés sur des modèles sensibles et de permettre les contributions françaises aux évaluations internationales (par exemple AI Safety Institutes, bureau de l'IA de la Commission européenne).

CONTEXTE

L'évaluation des systèmes d'IA nécessite une infrastructure logicielle et matérielle adaptée aux besoins de l'INESIA, tels que définis dans la présente feuille de route. Ces besoins pourraient inclure les cas d'usage suivants sans s'y restreindre : évaluation de modèles d'IA à usage général et des agents autonomes, participation aux travaux des AISI si opportun, conception et mise à jour régulière des composants de base (*datasets*, outils d'évaluation, scripts de réplication) nécessaires au maintien de *leaderboards* fiables. La plateforme devra en particulier répondre aux spécificités des évaluations de modèles qui requièrent un fort niveau de confidentialité, qu'ils soient non publics ou à haut risque. Elle pourra permettre l'hébergement de jeux d'évaluation privés nécessaires à la bonne conduite des activités.

Comme évoqué et en tant que de besoin, des articulations avec les autres projets, notamment Projet n°6., Projet n°7., Projet n°9, Projet n°11 seront considérés.

MÉTHODOLOGIE

Il convient dans un premier temps de préciser les cas d'usage dans le contexte de l'INESIA ainsi que les pratiques actuelles des acteurs dans les différents projets. Ensuite, une évaluation des frameworks open-source pour l'évaluation de modèles génératifs sera mise en place ainsi que les outils nécessaires à l'interfaçage avec les expérimentateurs, à l'orchestration et au suivi des expériences.

À l'aune d'une comparaison des besoins identifiés et des capacités existantes recensées, cette étude pourra donner lieu à la production d'un premier cahier de charges, spécifiant la plateforme visée pour répondre aux besoins de l'INESIA. À ce titre, la possibilité de contribuer à l'écosystème open source sera explorée, si un lien avec les activités de l'INESIA s'avère pertinent. Les travaux dans ce projet se structureraient autour de trois éléments :

- l'interface avec les utilisateurs (expérimentateurs) ;
- l'orchestration et suivi des expériences, avec l'ajout des briques logicielles personnalisées ;
- les moyens de calcul mis à disposition.

LIVRABLES

- Détermination des besoins de l'INESIA, recensement des capacités disponibles et selon l'analyse proposée et les ressources disponibles, conception d'une infrastructure logicielle et matérielle pour le lancement d'expérimentations avec des modèles d'IA et des agents.

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN. Des partenaires extérieurs pourront être associés.

PROJET N°11.

ANIMATION SCIENTIFIQUE AUTOUR DES TRAVAUX DE L'INESIA

OBJECTIF

Structurer, animer et fédérer la communauté nationale voire internationale autour de la thématique de l'évaluation de l'IA. Accélérer la montée en maturité des méthodologies d'évaluation, favoriser le partage des meilleures pratiques, stimuler la confrontation des approches, renforcer la visibilité et l'attractivité de la recherche française dans ce domaine stratégique. Faciliter la veille sur les sujets scientifiques d'intérêt pour l'INESIA.

CONTEXTE

L'évaluation de l'IA fait face à des verrous méthodologiques majeurs : absence de consensus sur les métriques, diversité des cas d'usage, incertitude sur les propriétés à mesurer. L'INESIA doit ainsi fédérer les expertises académiques, industrielles et institutionnelles, afin de discuter de l'état de l'art.

MÉTHODOLOGIE

L'INESIA organisera des journées scientifiques annuelles pour partager les différentes avancées méthodologiques ou technologiques sur l'évaluation des IA. Pour cela elle effectuera une veille méthodologique afin d'identifier les dernières avancées dans ce domaine. Ces avancées seront listées dans un document de veille bibliographique et les plus importants seront partagés aux différents partenaires d'INESIA. Les résultats des challenges pourront également y être présentés.
Des appels à communications pourront être ouverts et couvriront un large spectre : fondements méthodologiques, retours d'expérience, outils open source, benchmarks sectoriels, etc.
Par ailleurs, des activités communes de veille scientifique pourront être menées, et donner lieu à des points réguliers sous un format à préciser.
Par nature, ce projet est amené à s'interfacer avec les autres projets, afin de mettre en valeur leurs productions.

LIVRABLES

- Actes de la conférence (communications, posters, recommandations, etc.)

PARTIES PRENANTES AU PROJET

ANSSI, Inria, LNE et PEReN. Le milieu académique, certaines start-ups, industries, plateformes de challenges pourront être associés.