GOUVERNEMENT
Liberté
Égalité
Fraternité

COMMUNIQUE DE PRESSE

Paris, le 12/02/2025

Sommet pour l'action sur l'IA : Assurer le développement d'une IA de confiance, sûre et sécurisée, au service de tous

L'intelligence artificielle (IA) transforme profondément notre société, ouvrant des perspectives inédites dans de nombreux domaines. Pour que cette révolution technologique bénéficie à tous**, il est crucial de garantir un développement responsable et éthique, plaçant la confiance au cœur de son essor**. L'IA soulève en effet d'importantes questions en matière de sécurité et de sûreté, les sommets de Bletchley Park (Royaume-Uni – novembre 2023) et Séoul (Corée du Sud – mai 2024) en ont clairement fait la démonstration. Qu'il s'agisse d'anticipation des risques les plus extrêmes ou de dérives déjà visibles, **un traitement résolument ambitieux des facteurs de confiance en l'IA est par conséquent indispensable, à l'échelle internationale**.

Les progrès technologiques liés à l'IA apportent également des possibilités exceptionnelles dans le champ de la sécurité. C'est dans cette optique que le Sommet pour l'action sur l'IA s'est engagé concrètement pour promouvoir une IA sûre et sécurisée, notamment en mettant à la disposition de tous des outils de remédiation face à ces risques.

Pour que l'IA tienne toutes ses promesses, une approche collaborative est essentielle. Le Sommet pour l'action sur l'IA appelle à la mobilisation des acteurs publics, privés et académiques pour construire ensemble un écosystème de confiance autour de l'IA.

Cette approche mondiale repose sur trois piliers : la science, les solutions et les standards. En s'appuyant sur un consensus scientifique international robuste en matière d'IA, **le temps est venu de développer des solutions techniques ouvertes et accessibles à tous** et de créer des normes internationales communes reconnues par l'ensemble de l'écosystème, afin d'éviter la fragmentation et pour encourager la convergence à tous les niveaux.

Dans le cadre du Sommet pour l'action sur l'IA, le ministère de l'Europe et des affaires étrangères et le Secrétariat général de la défense et de la sécurité nationale ont appuyé l'action de l'envoyé thématique, Guillaume Poupard, afin de fédérer un écosystème d'acteurs, à la fois à l'échelle nationale et internationale, mobilisés pour renforcer la sûreté et la sécurité de l'IA.

ÉVALUER ET REPONDRE AUX ENJEUX DE SECURITE ET DE SURETE DE L'INTELLIGENCE ARTIFICIELLE

AI Safety Report.

Dans la continuité des Sommets de Bletchley Park et de Séoul, les travaux relatifs aux risques des modèles d'IA avancés se sont poursuivis sous l'égide de Yoshua Bengio, qui s'est entouré d'un panel d'experts internationaux reconnus. Les conclusions de ce rapport, présentées officiellement lors du Sommet, ont permis d'ouvrir la voie à des discussions nourries entre experts internationaux lors des journées scientifiques et tout au long du Sommet. Ce rapport entend décrire la façon dont les modèles d'IA à usage général fonctionnent, à identifier les risques qu'ils peuvent poser et à s'intéresser aux moyens d'y faire face.

Le rapport et ses conclusions continueront d'être enrichis dans le temps pour tenir compte des dernières évolutions, sous secrétariat britannique.

> International AI Safety Report 2025

Création de l'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA).

Le 31 janvier Clara Chappaz, ministre déléguée chargée de l'Intelligence artificielle et du Numérique, a annoncé le lancement d'un **Institut national consacré à l'évaluation et à la sécurité de l'IA (INESIA)**. Piloté par le secrétariat général de la défense et de la sécurité nationale (SGDSN), au nom du Premier ministre, et par la direction générale des entreprises (DGE) du ministère de l'économie, des finances et de la souveraineté industrielle et numérique, cet institut permettra de fédérer un écosystème d'acteurs nationaux de premier rang.

Face aux défis posés par l'IA, la science doit objectiver les risques pour mieux les anticiper. Par la recherche, l'évaluation et la coopération internationale, il est ainsi possible de promouvoir une IA responsable, centrée sur l'humain et le bien commun.

Cela passe aussi par le **développement d'outils techniques fiables, capables de garantir la sécurité et de prévenir les risques liés à l'IA**.

C'est dans ce contexte et en déclinaison de la Stratégie nationale sur l'IA que le gouvernement français a décidé la création de l'INESIA.

Le SGDSN et le DGE, qui en assureront le pilotage, auront la responsabilité de fédérer un écosystème dynamique de chercheurs et d'ingénieurs – issus en premier lieu de l'ANSSI, de l'INRIA, du LNE et du PEReN – qui, dans une logique d'intérêt général et de collaboration, se concentrera sur **l'analyse des risques systémiques dans le champs de la sécurité nationale, le soutien à la mise en œuvre de la régulation de l'IA, et l'évaluation de la performance et de la sûreté de fonctionnement des modèles d'IA**.

Les travaux sur l'évaluation de l'IA sont, à l'échelle nationale, soutenus par France 2030, dans un contexte où les enjeux de sécurité et de fiabilité sont un levier majeur de compétitivité pour les acteurs français de l'IA, qui devront atteindre les standards en la matière.

L'INESIA rejoint le groupe des « *AI Safety Institutes* » lancés lors du sommet de Bletchley Park, aujourd'hui présents dans une dizaine de pays et constitués en réseau à San Francisco en novembre 2024. Les **premiers résultats d'expérimentations conjointes**, très prometteurs, ont été publiés lors du Sommet.

Ces essais témoignent de l'ambition de réseau d'évaluer des modèles d'IA en s'appuyant sur la science et d'élaborer des bonnes pratiques pour tester les systèmes d'IA avancés. Singapour, le Japon et le Royaume-Uni ont conduit ces évaluations conjointes, à la fois afin de tester la robustesse cyber de systèmes d'IA et les réponses fournies par les modèles en fonction de la langue de l'utilisateur.

Cette nouvelle collaboration internationale est fondamentale pour la conduite de ces évaluations afin de s'assurer de leur robustesse, leur précision et la prise en compte des spécificités linguistiques. Cette

coopération est d'autant plus cruciale lorsque l'évaluation des modèles porte sur des questions de défense et de sécurité nationale. Dans le cadre de cette expérimentation, la France a fourni le jeu de données ayant permis l'une des deux évaluations, portant sur la cybersécurité des modèles.

> Expérimentations conjointes

Lancement du *leaderboard* des modèles de langage pour le français.

Afin de contribuer à comprendre et évaluer les technologies d'IA avec rigueur, transparence et indépendance, la Coordination nationale pour l'IA, en collaboration avec le ministère de l'Éducation nationale, l'Inria, le LNE et GENCI, a uni ses forces avec la *startup* Hugging Face pour construire un ***leaderboard* de référence dédié aux modèles de langage**.

Ce leaderboard compare, selon des critères objectifs et transparents, les modèles de langage adaptés à la langue française, sur des jeux de données en français, adaptés aux spécificités culturelles de la francophonie.

Son objectif est clair : **améliorer les performances en langue française des meilleurs modèles de langage indépendamment de leur lieu de développement**. Cela garantira à toutes les entreprises et institutions francophones un accès aux technologies les plus avancées, optimisées pour leur langue de travail et leurs usages spécifiques.

> Leaderboard des modèles de langage français

L'IA DANS LE CHAMP DES INGERENCES NUMERIQUES ETRANGERES, DE LA CYBERSECURITE ET DE LA PROTECTION DES DONNEES

Lutter contre les manipulations de l'information.

L'essor de l'IA, combinée aux mécanismes algorithmiques des réseaux sociaux, interroge sur ses conséquences en matière d'intégrité de l'information. En particulier, cette combinaison technologique fait peser le **risque d'une altération de la perception de la réalité par les citoyens**, susceptible de produire des effets négatifs durables sur le fonctionnement de la démocratie, fondée sur la notion de confiance. Aussi, dans le cadre du Sommet, un ensemble d'acteurs, français et internationaux, provenant aussi bien des milieux publics que privés, institutionnels qu'académiques, ont mis en commun leurs expertises et savoir-faire pour **explorer les défis et opportunités de l'IA pour la protection de l'intégrité de l'information et la lutte contre les manipulations de l'information**.
Afin de rassembler toutes les parties prenantes autour de ces enjeux, VIGINUM et l'OCDE ont organisé, avec le concours du Ministère de la Culture, un événement officiel en marge du Sommet, le 11 février.

En accompagnement de cet évènement, VIGINUM a publié un **rapport sur les défis et possibilités de l'IA dans la lutte contre les manipulations de l'information**. Intégrant la perspective de partenaires internationaux, ce rapport est construit autour d'une idée centrale forte : si les technologies d'IA sont susceptibles d'accroître la menace liée aux ingérences numériques étrangères, elles offrent également de formidables possibilités de renforcer nos défenses.

> Rapport VIGINUM

Afin de concrétiser ces possibilités, VIGINUM a mis à disposition un **outil open-source, « D3lta », capable d'identifier des tactiques de manipulations textuelles à grande-échelle** grâce à l'IA.

> D3lta sur GitHub

Un second outil, conçu par le Pôle d'expertise de la régulation numérique, le PEReN, en collaboration avec VIGINUM, vise, *via* un méta-détecteur qui permet d'évaluer de manière standardisée la performance de différents détecteurs de contenus artificiels, à **faciliter la détection de ces contenus en**

fonction de leur nature et de leur typologie. À terme cet outil pourra, en combinant les performances des détecteurs qui le composent, permettre de détecter des contenus artificiels jusqu'alors non identifiés comme tels.

> Outil de métadection - PEReN

S'agissant des acteurs de l'information, une **étude centrée sur les applications pratiques des outils d'IA dans les rédactions**, réalisée par des journalistes pour leurs confrères, a été publiée à l'occasion du Sommet.

Par ailleurs, la création d'un **réseau mondial de *fact-checking*** a été annoncée. Enfin, une déclaration commune des médias est venue détailler les mesures spécifiques visant à préserver le contrôle et l'intégrité d'informations fiables et professionnelles, au nom de la démocratie.

Renforcer la cyber-résilience.

L'utilisation et le déploiement des technologies d'intelligence artificielle à grande échelle soulèvent des défis importants en termes de cybersécurité sur un triple plan :

- la **sécurisation des systèmes d'IA** à travers l'élaboration de schémas d'évaluation et l'accompagnement des acteurs dans leur projet d'IA,
- le **développement d'outils de cybersécurité intégrant de l'intelligence artificielle** afin de renforcer leurs capacités, et enfin
- le renforcement de la **connaissance sur l'utilisation cyber-malveillante** des technologies d'IA.

En lien étroit avec ses partenaires nationaux et internationaux, l'ANSSI a travaillé à l'identification et à la bonne compréhension des risques cybersécuritaires des systèmes d'IA, afin d'en sécuriser et favoriser le déploiement. Une **analyse de risques menée internationalement** a été présentée le 7 février à l'occasion des journées scientifiques organisées par l'Institut Polytechnique de Paris. Ce travail de recherche met en exergue la nécessité d'adaptation aux défis cybersécuritaires de l'intelligence artificielle.

Sur le fondement de cette analyse, un **exercice de crise** a été organisé le 11 février, rassemblant plus de 200 participants issus du monde de l'IA et de celui de la cybersécurité. Son objectif était de favoriser le dialogue et, au travers de scénarios opérationnels, d'identifier les mesures de défense et de gouvernance pour renforcer la confiance dans les systèmes d'information intégrant de l'IA. Un **retour d'expérience** mettant en exergue les enseignements tirés de cet exercice sera publié à l'issue du Sommet afin de partager plus largement la **culture de la gestion de crise cybersécuritaire dans le champ de l'IA**.

> Analyse de risque cyber des systèmes d'intelligence artificielle

Protection de la vie privée.

Le développement des technologies d'IA est porteur d'enjeux essentiels en termes de vie privée. En effet, de très nombreux systèmes utilisent des données personnelles tant pour leur développement qu'au stade de leur utilisation. Les règles de protection des données offrent un cadre de gouvernance clair et protecteur favorisant la confiance. Cela est particulièrement important dans le contexte actuel de développement et de déploiement de l'IA, où le traitement des données devient extrêmement complexe, mobilise une chaîne de valeur très éclatée, opère à une très large échelle et emporte de grands risques d'opacité pour les individus.

Plusieurs événements se sont attachés à éclairer ces enjeux sous différents angles à l'occasion du Sommet et à y apporter des réponses, en particulier en ce qui concerne :

- La **protection des données**, et plus spécifiquement les enjeux autour du traitement de la voix et de l'image à l'heure de l'IA ;
- La **gouvernance**, à travers les questions de prise en compte dans un environnement technologique mondialisé des enjeux de vie privée au niveau mondial, en clarifiant les exigences aux parties prenantes, fluidifiant les échanges entre régulateurs et apportant de la sécurité aux individus.

La **mise en œuvre concrète de la régulation européenne**, à l'heure où de nombreux textes trouvent à s'appliquer : Règlement général sur la protection des données (RGPD), Règlement européen sur l'IA (RIA), Règlement sur les services numériques (RSN), *etc.*

GOUVERNANCE DE LA CONFIANCE

Afin de conjuguer innovation et développement sûr et éthique de l'intelligence artificielle, l'Union européenne s'est dotée en 2024 du **règlement européen sur l'IA**, premier cadre contraignant et complet au niveau mondial, permettant de mettre en œuvre des règles harmonisées sur le marché intérieur. C'est aujourd'hui le cadre de référence en France, permettant de placer la confiance au cœur des systèmes d'intelligence artificielle. Alors que les règles régissant le développement des modèles d'IA à usage général entreront en vigueur en 2025, la Commission européenne travaille actuellement, avec l'ensemble des acteurs du secteur, sur la **conception d'un code de bonnes pratiques, permettant de faciliter la bonne mise en œuvre de ces règles**. Le Sommet fut l'occasion de réaliser un point d'étape de ces travaux.

En parallèle, la France et ses partenaires internationaux se sont engagés dans la construction de **principes de confiance destinés à être adoptés sur une base volontaire par les entreprises à une plus grande échelle**, ciblant également les produits développés hors du marché européen. C'est ainsi qu'en marge du Sommet a été lancé le **Code de conduite issu du Processus d'Hiroshima**, négocié par le G7 depuis 2023 et visant à s'assurer que les organisations adhérentes, à savoir les développeurs issus des pays du Partenariat mondial pour l'intelligence artificielle (PMIA), appliquent un certain nombre de principes clefs en matière d'éthique et de sûreté de l'IA. En partenariat avec l'OCDE, un exercice de revue sera lancé cette année, permettant d'évaluer la bonne mise en œuvre du code de conduite par les entreprises. Les résultats en seront publiés régulièrement.

Enfin, d'autres initiatives impliquant les développeurs ont également été lancées lors de précédents Sommets – comme les *Frontier AI Safety Commitments* de Séoul – ou bien par d'autres organisations internationales (OCDE, ONU, UNESCO…). C'est dans ce cadre que le Sommet pour l'Action sur l'IA fut **l'occasion pour l'ensemble de ces parties prenantes de se réunir afin de présenter et comprendre ces dispositifs**, de permettre aux entreprises, aux universitaires et à la société civile d'y contribuer pleinement, mais aussi d'esquisser des convergences entre les différentes initiatives. La gouvernance de la confiance en matière d'intelligence artificielle a connu des progrès considérables ces dernières années, menant parfois à une multiplication voire une duplication des initiatives. **La philosophie du Sommet était ainsi de continuer à clarifier cette architecture, pour une gouvernance à la fois efficace et inclusive**.