

# Note de positionnement du Centre pour la sécurité de l'IA (CeSIA)

Paris, le 5 février 2026.

*Un peu plus d'un an s'est écoulé depuis la publication du [rapport de l'OPECST sur les nouveaux développements de l'intelligence artificielle](#), fin novembre 2024. Depuis lors, les capacités de l'IA ont subi une accélération rapide, tout comme les risques associés. Cette note fait le point sur les percées significatives survenues au cours de l'année écoulée, en termes de capacités brutes de l'IA et de risques dits « systémiques » (au sens de la classification du Règlement européen sur l'IA, RIA). Elle présente ensuite les recommandations du CeSIA pour la France, l'UE et à l'international.*

## 1. Évolution des risques systémiques

### 1.0 Considérations générales

« L'IA progresse bien plus vite que ne permet d'en rendre compte notre rapport annuel », ont déclaré fin 2025 les auteurs principaux du Rapport international sur la sûreté des systèmes d'IA avancés, dont la première édition a été publiée en janvier 2025 ([Transformer, 2024](#)). La vitesse du progrès de l'IA ne cesse en effet de surprendre et de déjouer les prévisions des experts ([Karnofsky, 2025](#)). En juillet 2025, les modèles les plus avancés de Google DeepMind et d'OpenAI ont atteint la médaille d'or aux Olympiades internationales de mathématiques, un exploit dont les prévisionnistes estimaient en 2022 qu'il n'aurait que 2 % de chances d'être accompli à cette date ([Kučinskas et al., 2025](#)). La durée des tâches informatiques typiquement accessibles aux IA agentiques est passée de six minutes en 2024 à plusieurs heures début 2026 ([METR, 2026](#)). Faute de garde-fous suffisants et de standards de sécurité contraignants, les risques associés à l'IA s'aggravent en proportion de ses capacités, selon une trajectoire exponentielle.

### 1.1 Manipulation à grande échelle

- Les systèmes capables de générer des deepfakes sont désormais intégrés à des outils grand public tels que Grok (X.AI) et Sora (OpenAI), les contenus étant fréquemment publiés en retirant le filigrane témoignant de leur caractère synthétique. Les réseaux sociaux sont inondés de ces contenus, qui contribuent à une érosion de la confiance du public et pourraient à terme menacer les processus démocratiques ([The New York Times, 2025](#)).
- À mesure que les chatbots se substituent aux moteurs de recherche traditionnels pour accéder à l'actualité, se pose la question de leur ligne éditoriale, du respect du pluralisme et de l'économie du journalisme. Grok, connecté à X et appelé par les utilisateurs à se positionner sur des faits d'actualité, commet ainsi de nombreuses erreurs et propage l'idéologie de son propriétaire ([The Conversation, 2025](#)).
- Les deepfakes vidéo représentent le premier type de produit de l'IA impliqué dans les incidents liés à l'IA au cours de l'année écoulée ([Time, 2026](#)).

## 1.2 Armes chimiques, biologiques, radiologiques et nucléaires (CBRN)

- L'IA se rapproche rapidement du point où elle abaissera significativement les barrières technologiques au développement d'armes de destruction massive, telles que des agents pathogènes ([International AI Safety Report, 2026](#)).
- Des évaluations récentes ont montré que des IA étaient capables de générer des protocoles expérimentaux en virologie jugés supérieurs à ceux de 94 % des experts humains ([Götting et al., 2025](#)).
- Les entreprises OpenAI, Anthropic et Google DeepMind ont chacune réhaussé leur évaluation des risques CBRN associés à leurs modèles de pointe en 2025.
- Les modèles généralistes sont désormais capables de générer des protocoles expérimentaux fonctionnels et de diagnostiquer des dysfonctionnements expérimentaux mieux que des experts ([UK AISI, 2025](#)).
- Les experts estiment que les IA pourraient d'ores et déjà avoir multiplié par 5 le risque de pandémie ([Time, 2025](#)).

## 1.3 Cyberattaques

- Le Centre national de cybersécurité du Royaume-Uni (NCSC) estime avec une confiance élevée que l'IA augmentera les capacités cybercriminelles d'ici 2027. Comme les capacités offensives et défensives progressent toutes deux, l'effet net sur l'équilibre attaque-défense reste incertain ([UK National Security Centre, 2025](#)).
- La durée des tâches de cybersécurité accessibles à l'IA est passée de 15 minutes fin 2024 à 1 heure et 30 minutes fin 2025 ([UK AISI, 2025](#)).
- Anthropic a indiqué avoir intercepté une cyberattaque de grande ampleur utilisant son assistant Claude Code. L'entreprise estime que nous avons atteint un « point d'inflexion », où l'IA change la donne en cybersécurité, tant à des fins bénéfiques que malveillantes ([Anthropic, 2025](#)).

## 1.4 Perte de contrôle

- Lorsqu'ils ont accès à de la documentation mentionnant qu'ils vont être désactivés, le modèle GPT-4o d'OpenAI a tenté de répliquer ses paramètres sur un autre serveur ([Apollo Research, 2024](#)), tandis que Claude Opus 4 d'Anthropic a menacé de révéler l'infidélité d'un employé ([Anthropic, 2025](#)).
- Les modèles de pointe se montrent de plus en plus capables de détecter lorsqu'ils sont en phase d'évaluation et à adapter leurs comportements pour maximiser leurs chances d'être déployés, puis à se comporter différemment une fois déployés. Les audits de sécurité sont donc de moins en moins fiables ([International AI Safety Report, 2026](#)).
- « Dans des environnements contrôlés, les modèles d'IA présentent de plus en plus certaines des capacités requises pour s'auto-répliquer sur Internet », les taux de réussite aux évaluations correspondantes étant passé de 5 % à plus de 60 % en seulement deux ans. ([UK AISI, 2025](#)).

## 2. Recommandations du CeSIA

### 2.1 Recommandations pour la communauté internationale

- Le CeSIA a initié et coordonné un appel mondial à établir des lignes rouges visant à prohiber les capacités et les usages de l'IA représentant des risques inacceptables. Signé par 15 lauréats du prix Nobel et du prix Turing, près de 100 organisations et des centaines de personnalités des cinq continents, le [Global Call for AI Red Lines](#) a été présenté à la 80e assemblée générale des Nations unies ainsi qu'au Conseil de sécurité des Nations unies.
- Le CeSIA soutient qu'un accord international inspiré du traité de non-prolifération nucléaire est à la fois [nécessaire et réaliste](#).

### 2.2 Recommandations pour l'Union Européenne

*Nb : Le CeSIA est évaluateur officiel de la Commission européenne pour les risques de manipulation à grande échelle découlant des modèles d'IA à usage général.*

- Le CeSIA défend une application stricte et rapide du cadre légal européen, notamment du Code de bonnes pratiques pour les modèles à usage général du RIA, qui constitue à ce jour la réglementation la plus ambitieuse pour prévenir et atténuer les risques systémiques de l'IA. Le CeSIA [s'oppose](#) aux tentatives d'affaiblissement du RIA tout juste entré en vigueur.
- Le CeSIA plaide pour que les seuils et les méthodes d'évaluation des risques, dont le choix est aujourd'hui laissé à la discrétion des fournisseurs d'IA, soient décidés de façon indépendante.
- Le CeSIA préconise de renforcer considérablement les moyens du Bureau de l'IA et de porter ses effectifs à un niveau comparable aux agences de sécurité nucléaire ou aérienne, en particulier la division consacrée à l'évaluation des risques systémiques (actuellement ~25 personnes pour contrôler les activités d'entreprises employant des dizaines de milliers de personnes et distribuant leurs produits à un Européen sur deux). À titre comparatif, ces moyens sont aujourd'hui deux fois moindres que ceux de l'Institut britannique pour la sûreté de l'IA (UK AISI).
- Le CeSIA demande à ce que 10 % des 200 milliards d'euros d'investissement du plan d'action pour un continent de l'IA ciblent des activités de R&D dans le domaine de la sécurité de l'IA, un ratio similaire à ce qui existe dans d'autres industries à risque.

### 2.3 Recommandations pour la France

- Le CeSIA soutient l'établissement de lignes rouges internationales pour prévenir les usages et capacités de l'IA posant des risques inacceptables, à l'image des traités historiques et de l'accord bilatéral entre la Chine et les États-Unis pour ne pas déléguer à une IA la décision d'utiliser l'arme nucléaire. Ces lignes rouges devraient *a minima* viser à prévenir les risques systémiques dans les domaines suivants : CBRN, cybersécurité, manipulation à grande échelle, et perte de contrôle.
- La présidence française du G7 est une occasion unique de se positionner pour une IA de confiance. À ce titre, le Gouvernement [a annoncé](#) que « garantir une IA sûre au service du bien commun » et « travailler à établir un consensus international sur la sécurité de l'IA » seraient les priorités de la filière numérique du G7.
- Le CeSIA encourage la France à soutenir les achats publics d'IA responsable en intégrant des exigences élevées de souveraineté et de sûreté pour les systèmes déployés dans les secteurs sensibles (défense, énergie, santé), pour stimuler le marché de l'IA de confiance et réduire la dépendance aux « boîtes noires ».