

GARANTIES DE CONTRÔLE DE L'IA ET MÉCANISMES DE VÉRIFICATION

Implications et options
pour la France

RÉSUMÉ EXÉCUTIF

— CONTEXTE —

Le 7 avril 2026, Anthropic a annoncé le lancement restreint de **Claude Mythos**, un modèle d'IA dont les capacités offensives en cybersécurité ont été jugées trop dangereuses pour une diffusion publique. Lors de tests, le modèle s'est notamment révélé capable de **planifier et d'exécuter des séquences d'attaques de manière autonome**. Les prochaines générations de modèles, plus performants, sont déjà en cours d'entraînement. **Or aucune méthode connue ne permet aujourd'hui d'en garantir le contrôle et aucun mécanisme ne permet de vérifier que ceux qui les développent respectent leurs engagements**. C'est précisément dans cet interstice, face à l'avancée des performances de l'IA et en l'absence de garanties de contrôle, que **la France peut faire valoir un rôle singulier**. Celui de contribuer à bâtir les mécanismes de vérification sans lesquels aucune coordination internationale ne saurait être durable et crédible.

Tom David, Président Directeur Général
du General-Purpose AI Policy Lab

— QUESTIONS —

- Quelles sont les conséquences de la course à l'AGI entre les États-Unis et la Chine ?
- Quels dispositifs permettraient à des acteurs rivaux de coopérer en confiance pour un développement contrôlé et sécurisé de l'IA de frontière ?

— SYNTHÈSE —

La course à l'AGI entre les États-Unis et la Chine produit un nivellement par le bas des mesures de sécurité, au détriment de tous les acteurs internationaux. Cette course repose sur le présupposé que l'acteur qui développe la technologie en premier est capable de la contrôler. Or, il n'existe actuellement aucune méthode permettant de garantir le contrôle des modèles d'IA fondés sur l'apprentissage profond. En l'absence de telles garanties, le développement de systèmes d'IA de plus en plus performants, généraux et agentiques expose l'ensemble des acteurs à des problèmes de sécurité irréversibles à grande échelle.

Le développement de garanties de contrôle crédibles nécessite du temps et des investissements substantiels, mais la perception de dynamique de course dissuade chaque acteur de les prioriser unilatéralement. Seule une coordination internationale permettrait de réallouer massivement des ressources vers la mise au point de garanties de contrôle et vers les technologies d'IA ne posant pas de problèmes majeurs de sécurité, selon une logique de *développement technologique différentiel* (détaillée dans la note).

Cette coordination repose sur des engagements communs, eux-mêmes conditionnés à la mise en place de mécanismes de vérification crédibles. Ces mécanismes doivent satisfaire simultanément trois contraintes : empêcher les défections secrètes, protéger les informations industrielles et étatiques confidentielles, et être librement auditable pour s'assurer de l'absence de vulnérabilités exploitables.

Plusieurs solutions techniques sont envisageables, selon des horizons de déploiement distincts (voir Figure ci-dessous), en s'appuyant sur le paradigme technologique actuel nécessitant de grandes quantités de puissance de calcul. Ces stratégies peuvent être poursuivies en parallèle.

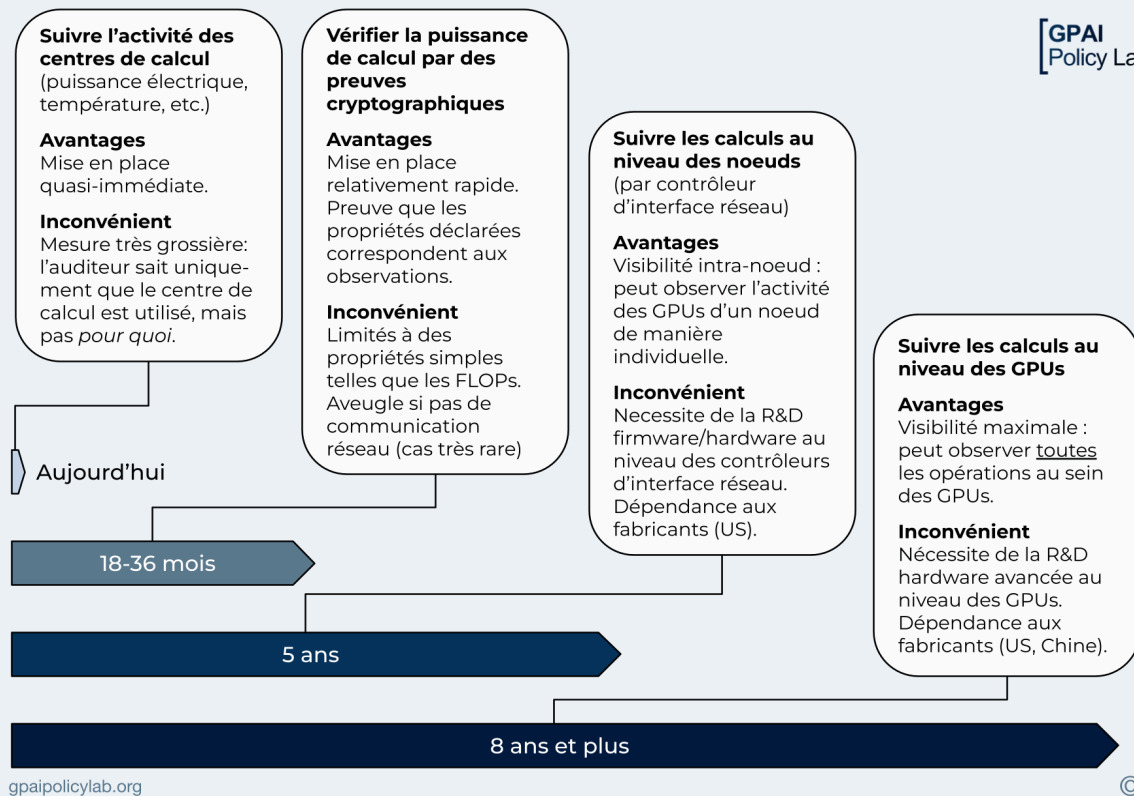


Figure A - Stratégies d'implémentations de mécanismes de vérification en fonction du temps nécessaire à leur déploiement.

Les mécanismes de vérification ne résolvent pas le problème fondamental de l'absence de garanties de contrôle des modèles d'IA de pointe, mais offrent des options réalistes permettant à des acteurs rivaux de se coordonner sur ces enjeux de sécurité.

— IMPLICATIONS POUR LES ACTEURS FRANÇAIS ET EUROPÉENS —

La course à l'AGI confronte la France et la communauté internationale à des problèmes de contrôle irréversibles.

Les mécanismes de vérification, en rendant possible la coordination internationale, ouvriraient une voie vers l'obtention de garanties de sécurité avant que des GPAI ne dépassent les capacités des experts humains dans des domaines critiques.

Cette situation offre à la France l'option de s'affirmer comme pionnière du développement des mécanismes de vérification.

La France possède une expertise reconnue en cryptographie et en sécurité matérielle, un écosystème industriel en microélectronique, et une tradition d'engagement dans les régimes internationaux de vérification. Alors qu'un écosystème de R&D sur la vérification émerge au Royaume-Uni et aux États-Unis, une fenêtre reste ouverte pour qu'un acteur comme la France se positionne sur ce segment technologique.

Garanties de contrôle de l'IA et mécanismes de vérification

Implications et options pour la France

I. La course à l'AGI et le nivellement des mesures de sécurité

*"Aujourd'hui, une nouvelle frontière de découvertes scientifiques s'ouvre à nous, définie par des technologies transformatrices telles que l'intelligence artificielle... **Les avancées dans ces domaines ont le potentiel de remodeler l'équilibre mondial des puissances, de donner naissance à des industries entièrement nouvelles et de révolutionner notre façon de vivre et de travailler. Alors que nos concurrents mondiaux se précipitent pour exploiter ces technologies, il est impératif pour la sécurité nationale des États-Unis d'atteindre et de maintenir une domination technologique mondiale incontestée et incontestable.**"*

Donald J. Trump, préambule du Plan d'action pour l'IA du gouvernement américain, juillet 2025¹.

En vertu des avantages stratégiques majeurs qu'elle pourrait fournir (percées militaires et cyber notamment), le développement d'une AGI (*Artificial General Intelligence*) est explicitement considéré comme un enjeu de sécurité nationale par certains États, en particulier la Chine² et les États-Unis³. Ces deux puissances apparaissent de plus comme les seuls compétiteurs crédibles au regard des ordres de grandeur requis en puissance de calcul, en énergie et en ressources, et de la mainmise américaine sur la chaîne de production.

Par ailleurs, les IA à usage général (GPAI) commencent déjà à être utilisées pour accélérer la R&D en IA, ce qui implique que le premier acteur à développer cette technologie peut espérer capitaliser sur celle-ci pour creuser l'écart avec ses rivaux⁴. Ce phénomène de rétroaction positive augmente sensiblement l'intérêt de ces pays non pas simplement à égaler leurs compétiteurs, mais à mener la course avec une longueur d'avance pour s'assurer un avantage décisif irréversible.

Ces positionnements stratégiques produisent une dynamique de course à l'AGI⁵ dont les conséquences prévisibles sont un nivellement par le bas (*race to the bottom*) des mesures de sécurité au profit d'un développement maximal des performances. Cette dynamique est documentée de longue date pour différentes courses technologiques et militaires, classiquement analysées à travers des modèles de théorie des jeux⁶ (cf. Encadré 1). La logique usuelle de course technologique repose toutefois sur le présupposé que l'acteur qui développe la technologie en premier en tire un avantage décisif, parce qu'il en maîtrise le fonctionnement. Or, dans le cas de l'IA à usage général, ce présupposé est mis en cause puisque le contrôle des modèles fondés sur l'apprentissage profond est, à ce jour, un problème de recherche ouvert auquel aucune solution n'a été apportée (détails dans la partie suivante).

¹ "America's AI Action Plan." *White House* (July 2025). www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

² Webster, et al. "Full Translation: China's 'New Generation Artificial Intelligence Development Plan'" *DigiChina* (2017).

digi.china.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017; Xi Jinping a qualifié l'IA de "technologie stratégique" dans le cadre de la 20e étude collective sur le renforcement du développement et de la supervision de l'intelligence artificielle. (2025) www.gov.cn/yaowen/liebiao/202504/content_7021072.htm

³ Chase and Marcellino. "Incentives for U.S.-China Conflict, Competition, and Cooperation Across Artificial General Intelligence's Five Hard National Security Problems." *RAND* (2025). www.rand.org/pubs/perspectives/PEA4189-1.html

⁴ Sytsma. "Decisive Economic Advantage. Modeling the Transition from Temporary First-Mover Leads to Economic Dominance in Artificial General Intelligence." *RAND* (2026). https://www.rand.org/pubs/research_reports/RRA4444-1.html; Davidson. "Could one country outgrow the rest of the world?" *Forethought* (2025). www.forethought.org/research/could-one-country-outgrow-the-rest-of-the-world

⁵ Ainsi, le Plan d'action pour l'IA du gouvernement américain est sous-titré sans ambiguïté "Winning the race".

⁶ Cimpanu, et al. "Artificial Intelligence Development Races in Heterogeneous Settings." *ArXiv* (2020). arxiv.org/abs/2012.15234; The et al. "Modelling and Influencing the AI Bidding War: A Research Agenda". *AIES* (2019). doi.org/10.1145/3306618.3314265.

Encadré 1 - La course à l'AGI sous l'angle de la théorie des jeux

La théorie des jeux modélise les situations dans lesquelles le résultat des décisions d'un acteur dépend des choix d'autres acteurs. Appliquée aux relations internationales, elle permet d'étudier les configurations pouvant mener à des dynamiques collectives conflictuelles ou, à l'inverse, inciter à la coopération. C'est notamment dans le contexte de la course aux armements nucléaires que la *RAND Corporation* a contribué à fonder le champ dans les années 1950-1960. Un rapport récent de RAND⁷ s'inscrit dans cette tradition en appliquant ce cadre à la compétition dans le développement de l'IA avancée.

Le modèle de RAND oppose deux acteurs qui doivent choisir entre une approche prudente (*Baseline*), et une approche qui poursuit la trajectoire actuelle de progrès rapide des IA de pointe (*Accelerated*). Ce choix traduit un arbitrage entre deux forces opposées. D'un côté, le premier acteur à développer un système d'IA avancé a une chance d'obtenir un avantage stratégique décisif. De l'autre, toute accélération augmente la probabilité d'un accident catastrophique touchant l'ensemble des acteurs.

La nature de la dynamique dépend alors du rapport entre l'avantage compétitif attendu de l'accélération et le risque associé. Lorsque le premier l'emporte sur le second, le jeu prend la forme d'un dilemme du prisonnier : chaque acteur a intérêt à accélérer quel que soit le choix de l'autre, et l'équilibre unique est l'accélération mutuelle. À l'inverse, lorsque les deux acteurs perçoivent le risque global comme suffisamment élevé, la structure bascule vers un jeu de coordination dans lequel deux équilibres coexistent : l'accélération mutuelle et la coopération mutuelle. La coopération devient alors stable et rationnelle, à condition que chaque acteur dispose d'assurances suffisantes sur le comportement de l'autre.

Le rapport de RAND permet d'identifier plusieurs facteurs qui favorisent ce basculement :

- Risque élevé de perte de contrôle. Plus la probabilité de ce scénario est élevée, plus les conséquences de l'accélération dépassent ses bénéfices stratégiques, et plus le jeu tend vers une structure de coordination.
- Incertitude sur l'issue de la course. Lorsque l'issue de la recherche reste imprévisible (risque de fuite technologique, diffusion rapide, etc.), accélérer ne garantit pas d'arriver premier. Le gain marginal de l'accélération diminue, ce qui réduit l'incitation à prendre des risques supplémentaires.
- Mécanismes de vérification crédibles. Les auteurs citent les mécanismes de vérification comme un levier d'intervention. En l'absence de vérification, même un jeu de coordination peut échouer, puisque chaque acteur, incapable de confirmer que l'autre respecte ses engagements, se rabat sur l'accélération.

Une analyse similaire⁸, réalisée dans le cadre du programme britannique *Safeguarded AI* de l'*Advanced Research and Invention Agency* (ARIA), confirme ce résultat. En faisant varier la durée pendant laquelle une trajectoire maîtrisée serait moins compétitive, ainsi que la part de valeur économique qu'elle préserverait pendant cette période, ce modèle montre que la structure stratégique se déplace d'un dilemme du prisonnier vers une situation favorable à la coopération lorsque le risque associé à la trajectoire par défaut est élevé.

⁷ Abraham, et al. "A Prisoner's Dilemma in the Race to Artificial General Intelligence." *RAND Corporation* (2025). https://www.rand.org/pubs/research_reports/RRA4245-1.html.

⁸ Dalrymple. "Safeguarded AI: constructing guaranteed safety." *Advanced Research and Invention Agency* (2024). <https://aria.org.uk/media/3nhjino4/aria-safeguarded-ai-programme-thesis-v1.pdf>.

Nos notes précédentes⁹, et la littérature associée, documentent le fait que, passé un seuil de performance, de généralité et d'agentivité, le développement de modèles d'IA à usage général pose des problèmes irréversibles à grande échelle, menaçant la sécurité nationale et internationale¹⁰. Les modèles de menace classiques passent par des usages malveillants (cyberattaques, bioterrorisme, etc.), mais les conséquences irréversibles les plus larges sont à attendre des scénarios de perte de contrôle¹¹. **L'absence de garanties de contrôle renverse les présupposés de suprématie technologique à l'origine de la course : un acteur qui développe un système qu'il ne peut pas contrôler non seulement ne dispose pas d'un avantage stratégique, mais il expose tous les acteurs à des problèmes de sécurité majeurs.** La Figure 1 ci-dessous illustre ce changement de perspective (voir Encadré 1 pour plus de détails sur le modèle de théorie des jeux sous-jacent).

Figure 1 - Intégration des enjeux de contrôle dans la dynamique de course à l'AGI. Les États-Unis et la Chine font face au même choix entre alimenter la course et opter pour un développement de l'IA maîtrisé. Cette situation est souvent analysée sous l'angle de la domination économique et militaire, mais cette perception est invalidée si les acteurs ne contrôlent pas les systèmes qu'ils développent¹².

GPAI Policy Lab		Chine	
		Course à l'AGI	Développement maîtrisé
USA	Course à l'AGI	→ Risques systémiques → Perte de contrôle	Perception de domination US → Risques systémiques → Perte de contrôle
	Développement maîtrisé	Perception de domination chinoise → Risques systémiques → Perte de contrôle	<u>Coordination</u> → Garanties de contrôle → Bénéfices de l'IA

Au vu du dilemme perçu par les grandes puissances de l'IA, aucune d'entre elles ne souhaite s'engager unilatéralement dans un développement davantage maîtrisé. **Cela implique que seul un effort de coordination permettrait d'atteindre une situation d'équilibre** propice à exploiter les bénéfices des GPAI sans s'exposer à des conséquences irréversibles. Cette coordination, reposant sur des engagements pris en commun, n'est à son tour possible que si des mécanismes de vérification du respect de ces engagements sont disponibles et mis en place par chacun des acteurs impliqués.

⁹ "Artificial General Intelligence (AGI) : Faisabilité technique et horizons temporels." GPAI Policy Lab (2025) ; "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI." GPAI Policy Lab (2025). <https://gpaipolicylab.org/note-4>.

¹⁰ Mitre, and Predd. "Artificial General Intelligence's Five Hard National Security Problems." RAND Corporation (2025). www.rand.org/pubs/perspectives/PEA3691-4.html ; General Purpose AI Policy Lab. "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle." GPAI Policy Lab (2025).

¹¹ Maier, et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." ArXiv, 2025, arxiv.org/abs/2510.02840 ; Le directeur du MI5 mentionnait en novembre 2025 « *while contending with today's threats, we also need to scope out the next frontier: potential future risks from non-human, autonomous AI systems which may evade human oversight and control.* » <https://www.mi5.gov.uk/director-general-sir-ken-mccallum-gives-threat-update>.

¹² Abraham, Kavner, and Moon. "A Prisoner's Dilemma in the Race to Artificial General Intelligence" RAND Corporation (2025). www.rand.org/pubs/research_reports/RRA4245-1.html ; Predd et al. "Finding Common Ground in AGI Strategy Debates" RAND Corporation (2026). <https://www.rand.org/pubs/perspectives/PEA4448-1.html>.

II. Assurer des garanties de contrôle des futures GPAI

Encadré 2 - Définitions

Garanties de contrôle : garanties que le comportement d'un système d'IA est contrôlable en toute circonstance et qu'il ne pose pas de problème grave de sécurité.

Exemple : preuve formelle de fonctionnement du modèle selon les spécifications des développeurs, dans toutes les situations possibles.

Engagements de conformité : engagement de la part des développeurs de modèles de ne pas mettre en œuvre de processus de développement et de déploiement non conforme aux règles établies en commun.

Exemple : engagement à ne pas utiliser une puissance de calcul supérieure à 10^{28} FLOPs pour entraîner un modèle d'IA à usage général.

Mécanisme de vérification : moyen technique permettant de vérifier la mise en application des engagements de conformité pris par un acteur.

Exemple : mécanisme de surveillance des opérations au niveau des serveurs permettant de monitorer le nombre de FLOPs utilisés durant un entraînement.

1. Les garanties de contrôle n'existent pas actuellement

Le comportement des modèles d'IA fondés sur l'apprentissage profond (*deep learning*) est en l'état impossible à prédire avec certitude : leurs mécanismes de prise de décision sont encodés dans des combinaisons de matrices de très haute dimension que l'on ne sait pas interpréter (empêchant de superviser leur cognition, et par conséquent de les contrôler avec haute confiance). C'est pour cette raison que ces systèmes sont qualifiés de « boîtes noires ».

Par ailleurs, les méthodes de développement de ces systèmes s'appuient exclusivement sur des approximations des objectifs souhaités par les développeurs¹³. Sous forte pression d'optimisation, ces méthodes produisent par défaut des divergences systématiques entre les objectifs attendus et ceux internalisés par le système d'IA (*Goodhart failure modes*¹⁴). **Cela implique qu'il n'existe aucune méthode connue permettant d'encoder des règles strictes au sein des modèles d'IA basés sur l'apprentissage profond et donc aucune garantie de contrôle.**¹⁵

2. Développer ces garanties nécessite du temps et des investissements

Pour les raisons détaillées précédemment, il ne sera probablement pas possible d'obtenir des garanties de contrôle dans le paradigme actuel du développement de l'IA. Or, aucune autre approche en IA n'est actuellement capable de fournir des performances comparables à l'apprentissage profond. Certains de ces paradigmes alternatifs visent explicitement à fournir des garanties de contrôle (approches *safe-by-design*¹⁶), mais le niveau d'effort et de ressources investis

¹³ Premièrement, les développeurs ne peuvent pas entièrement spécifier en lignes de code les objectifs à poursuivre et contraintes à respecter ; l'utilisation d'objectifs simplifiés induit donc une *misspecification*. Deuxièmement, quand les développeurs se basent sur des exemples pour inculquer des objectifs à l'entraînement des modèles, la distribution des données utilisées ne peut jamais être totalement représentative du monde réel, menant à des problèmes de généralisation.

¹⁴ Maier, et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." *ArXiv* (2025). arxiv.org/abs/2510.02840.

¹⁵ Une des démonstrations les plus flagrantes de cette absence de garantie de contrôle sur le comportement des modèles actuels est leur grande vulnérabilité aux instructions malveillantes (*jailbreaking*).

¹⁶ Dalrymple, et al. "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems." *ArXiv* (2024). arxiv.org/abs/2405.06624.

dans leur développement est aujourd'hui extrêmement faible en comparaison des investissements dans le développement des IA de frontière¹⁷.

La dynamique actuelle ne permet pas d'envisager que des solutions satisfaisantes aux questions de contrôle puissent être trouvées et déployées à l'échelle avant d'avoir atteint des seuils de performances exposant à des enjeux sérieux de perte de contrôle¹⁸.

3. La coordination permet de développer des garanties de contrôle

Le constat est donc le suivant : **développer des garanties de contrôle robustes nécessite de libérer du temps et des ressources, mais, tant que les problèmes de contrôle ne rentrent pas dans le raisonnement, aucune grande puissance de l'IA percevant un risque de prendre du retard dans la course à l'AGI ne souhaitera prioriser ces investissements.**

Par conséquent, seule une coordination au niveau international autoriserait les acteurs à accepter de modifier la dynamique de course, et à réallouer massivement des ressources vers des approches de contrôle et de sécurité. Cette stratégie est celle du *développement technologique différentiel*, qui consiste à investir en priorité dans des technologies de sécurité dites "défensives" qui permettront une fois maîtrisées de développer des technologies initialement trop à risque (cf. Encadré 3).

Encadré 3 - Développement technologique différentiel de l'IA

La stratégie du développement technologique différentiel permet de poursuivre le développement général de l'IA, à condition d'allouer significativement davantage de ressources à la mise au point de garanties de contrôle qu'à l'accélération de capacités des GPAI de pointe, tant que de telles garanties n'ont pas été établies.

À l'heure actuelle, alors qu'aucune garantie de contrôle n'existe, **seule une fraction des ressources des développeurs de modèles de frontière est allouée à résoudre ce problème**¹⁹. Un développement technologique différentiel impliquerait de rééquilibrer temporairement ce ratio.

Cette stratégie est ainsi compatible avec la poursuite d'investissements pour extraire les bénéfices des systèmes d'IA pouvant être déployés sans problèmes de sécurité majeurs dès aujourd'hui dans divers secteurs de l'économie (en particulier les IA *étroites* et *outils*).

Au vu de l'absence de garanties face aux conséquences irréversibles des scénarios de perte de contrôle, tous les acteurs internationaux ont donc intérêt à la mise en place d'une coordination internationale autour d'une stratégie globale de développement technologique différentiel. Du point de vue des intérêts français, le développement de futurs systèmes d'IA à usage général avancés par des acteurs tiers ne peut être envisagé que dans un cadre comportant des garanties strictes pour la sécurité. Les conséquences de la course à l'AGI sont du ressort de la communauté internationale dans son ensemble et non pas uniquement de ses seuls acteurs.

¹⁷ Le *Emerging Technology Observatory* estime que la sécurité de l'IA ne représente que 2% de la production de recherche en IA. Les approches *safe-by-design* ne représentent à leur tour qu'une infime partie de la recherche en sécurité de l'IA. "Still a drop in the bucket: new data on global AI safety research." ETO (2025). <https://eto.tech/blog/still-drop-bucket-ai-safety-research>.

¹⁸ "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI." GPAI Policy Lab (2025). <https://gpaipolicylab.org/note-4>.

¹⁹ En juillet 2023, OpenAI s'engageait à consacrer 20% de ses ressources aux questions de sécurité pour les 5 prochaines années. Moins d'un an après, l'équipe chargée de cette mission était dissoute. Son directeur, Jan Leike, a publiquement indiqué que les ressources promises ne leur étaient pas allouées et qu'il ne leur était par conséquent pas possible de faire des progrès sur les questions de sécurité et de contrôle dans ces conditions.

III. Des mécanismes de vérification sont nécessaires afin d'assurer le respect des engagements de conformité, et donc la coordination

La mise en place de mécanismes de vérification est une composante clé pour assurer à la fois des garanties de sécurité crédibles et la viabilité des efforts de coordination. **Chacun des acteurs, qu'il participe ou non à la course, doit avoir les moyens de s'assurer que les autres respectent les engagements collectifs de conformité**, sans quoi la confiance nécessaire pour faire évoluer la dynamique actuelle ne peut être établie. De tels mécanismes ont déjà été mis en place avec succès par le passé, sur d'autres technologies critiques (voir Encadré 4).

Encadré 4 - Exemples historiques de mécanismes de vérification

De nombreux accords internationaux ont impliqué, ou impliquent encore, la mise en place de mécanismes de vérification pour s'assurer du respect des engagements mutuels. Ces mécanismes peuvent s'appliquer par le biais de Moyens Techniques Nationaux (imagerie satellite) ou par une organisation spécifique chargée de cette mission (AIEA, Organisation pour l'interdiction des armes chimiques).

- **Le Traité sur la non-prolifération des armes nucléaires (TNP) (1970)** impose la déclaration des installations à l'AIEA et des mécanismes de vérification sur site continus, tels que l'installation de caméras de surveillance 24h/24, de capteurs de radiation et de scellés inviolables.
- **Les Traités SALT I (1972), ABM (1972), INF (1987), START I (1991) et New START (2010)** imposent la non-interférence avec les Moyens Techniques Nationaux (NTM) de vérification, garantissant le droit d'utiliser l'imagerie satellite et la télédétection pour surveiller les pays signataires.
- **Le Traité sur les forces nucléaires à portée intermédiaire (INF) (1987)** impose une surveillance continue du périmètre et des accès aux usines de production, incluant le droit de peser et de passer aux rayons X les véhicules sortants.
- **La Convention sur l'interdiction des armes chimiques (CIAC) (1997)** impose la déclaration exhaustive et l'inspection des installations de production, ainsi que des "inspections par mise en demeure" (*challenge inspections*) permettant à un État d'exiger l'accès à un site suspect d'un autre signataire dans un délai très court.

On se concentrera ici sur les mécanismes de vérification implémentés au niveau des centres de calcul utilisés par les entreprises d'IA. L'accès à de grands centres de calcul est aujourd'hui une condition *sine qua non* au développement des modèles d'IA de pointe. Cette dépendance à la puissance de calcul est donc considérée comme le levier le plus fiable pour mettre en place des moyens de vérifier les engagements de conformité. Ce principe fait écho au domaine nucléaire, où la vérification des stocks et flux d'uranium / plutonium / thorium par l'AIEA est un mécanisme clé pour garantir les engagements de non-prolifération.

Cette analogie repose toutefois sur la poursuite du paradigme technologique actuel de développement de l'IA. Si un changement radical de l'architecture des modèles venait à réduire de plusieurs ordres de grandeur la puissance de calcul nécessaire pour entraîner des modèles de pointe, les seuils et les dispositifs qui se basent dessus perdraient une partie de leur pertinence. Néanmoins, le paradigme actuel domine depuis plusieurs années le développement de l'IA et a mené aux avancées que l'on observe aujourd'hui, ce qui explique pourquoi les options présentées dans cette note présupposent la poursuite de cette trajectoire. Enfin, comme dans le domaine nucléaire, **l'utilité de ces mécanismes de vérification repose entièrement sur leur crédibilité**, laquelle est

conditionnée par leurs propriétés de résistance aux contournements, de confidentialité et de fiabilité, que nous allons maintenant détailler.

1. Les dispositifs de vérification doivent rendre impossibles les défections secrètes

Dans le contexte de très forte compétition internationale qu'est la course actuelle à l'AGI, une stratégie de *défection secrète* peut être considérée (à tort) comme avantageuse par certains acteurs.

> *Exemple : En l'absence de compréhension des enjeux liés aux scénarios de perte de contrôle, les USA et la Chine pourraient, à tort, chercher à capitaliser sur des engagements internationaux dans l'espoir que ceux-ci ralentissent leur rival, tout en poursuivant en parallèle un programme de développement secret violant ces accords afin de creuser l'écart technologique.*

Par conséquent, les dispositifs de vérification doivent être, soit idéalement en mesure de **résister à des tentatives de contournement de la part d'acteurs disposant de moyens étatiques**, soit a minima en mesure de rendre évident tout contournement (ou tentative de contournement).

> *Exemple : L'AIEA s'appuie sur des systèmes de scellés et de caméras de surveillance pour s'assurer du respect des engagements relatifs à la non-prolifération nucléaire. Il est possible pour un acteur de couper les caméras ou de briser les scellés mais cela rendrait la fraude évidente. La dissuasion repose, dans ce cas, non pas sur le fait de rendre la défection impossible mais sur le fait de la rendre manifeste.*

2. Les dispositifs de vérification ne doivent pas divulguer d'informations confidentielles

Le déploiement de dispositifs de vérification au sein des centres de calcul peut impliquer un accès à certaines informations relevant du secret industriel (architecture et hyperparamètres, données d'entraînement, procédure d'entraînement, poids des modèles). Dans ce cadre, le monitoring des activités doit garantir que ces informations soient strictement protégées.

> *Exemple : Des cas de fuites d'informations propriétaires de ce type en IA sont déjà avérés²⁰. Par conséquent, l'élargissement de l'accès à de telles informations à des tiers (auditeurs) augmente sensiblement les risques de divulgation d'informations confidentielles dans le contexte de compétition actuel.*

Afin de réduire ces risques, **les données confidentielles doivent être analysées au sein des centres de calcul sans moyens d'exfiltration**.

> *Exemple : Les dispositifs de vérification doivent être conçus de sorte à être incapables de transférer hors des centres de calcul des informations brutes, donc exclusivement des informations agrégées utiles mais non critiques (métriques d'intérêt, signatures issues de fonctions de hachage).*

3. Les dispositifs de vérification doivent être fiables et transparents

Le déploiement de dispositifs de vérification peut également introduire des risques d'exploitation de vulnérabilités à des fins d'espionnage industriel (vol de propriété intellectuelle) ou de sabotage.

> *Exemple : Un acteur malveillant (État ou organisation) pourrait exploiter des vulnérabilités pour faire passer un acteur respectant les engagements pour un tricheur afin de déclencher des réponses internationales indues à son égard.*

Par conséquent, **les dispositifs doivent être dépourvus de vulnérabilités et librement auditable** (afin que chaque acteur puisse s'en assurer).

²⁰ "Former Google Engineer Found Guilty of Economic Espionage and Theft of Confidential AI Technology." Department of Justice. www.justice.gov/opa/pr/former-google-engineer-found-guilty-economic-espionage-and-theft-confidential-ai-technology.

> Exemple : La méthode la plus efficace consisterait à co-construire des solutions open source (aussi bien au niveau logiciel que des composants électroniques) par le biais de projets de collaboration internationale. Cette approche permet à la fois de découvrir et corriger les vulnérabilités en amont, mais également de construire la confiance dans ces dispositifs de manière multilatérale.²¹

IV. Différentes options sont envisageables pour mettre en place des mécanismes de vérification

La vitesse de progrès des modèles de frontière implique de devoir arbitrer entre des solutions longues à développer mais qui permettraient des vérifications avancées et flexibles²², et des solutions rapides à mettre en production mais plus limitées en termes de métriques observables (voir Figure 2).

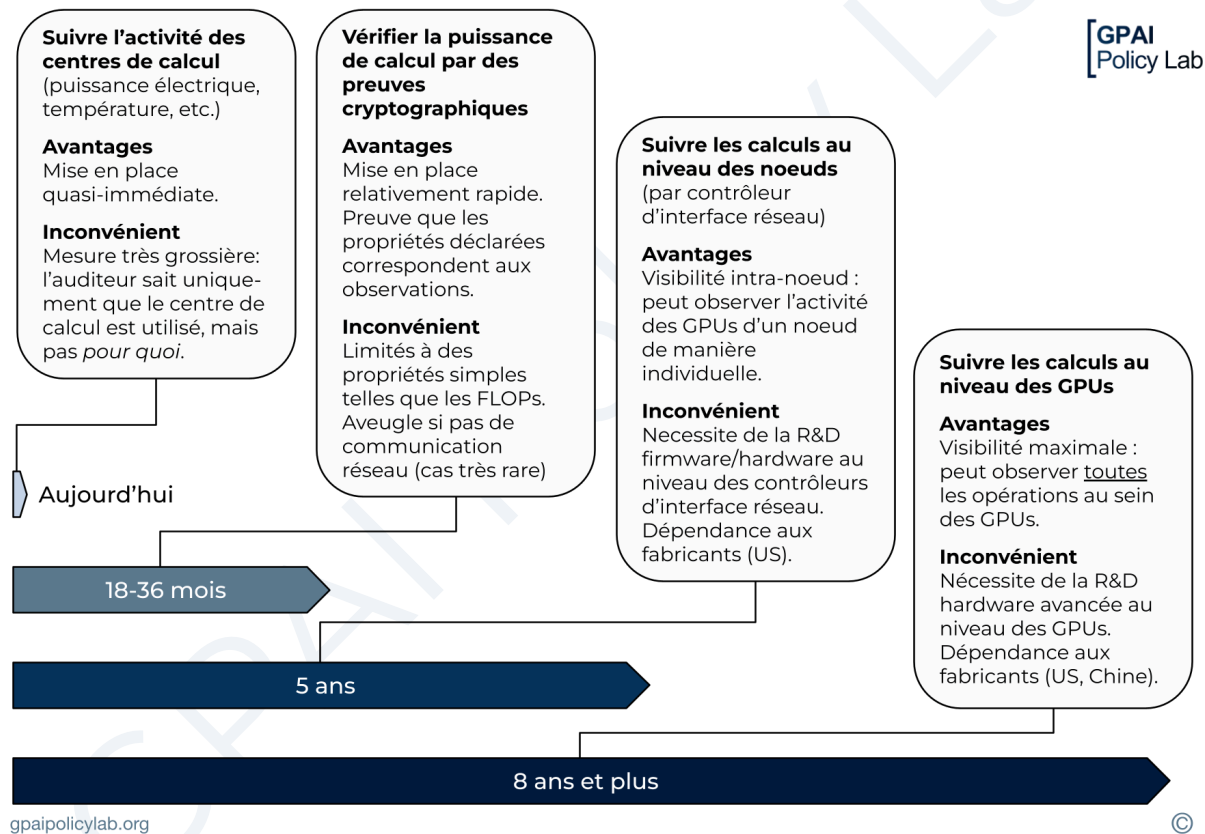


Figure 2 - Stratégies d'implémentations de mécanismes de vérification en fonction du temps nécessaire à leur déploiement.

Notons que ces stratégies peuvent être poursuivies en parallèle, en déployant très rapidement des solutions minimales respectant suffisamment toutes les contraintes critiques (robustesse, confidentialité, transparence et fiabilité), tout en investissant dans des solutions nécessitant davantage de recherche et développement, mais qui permettent après déploiement de s'adapter rapidement à l'évolution des méthodes d'entraînement d'IA et des choix de gouvernance. **Enfin, ces mécanismes visent à rendre possible la coordination, mais ils ne se substituent pas à l'effort de recherche sur les garanties de contrôle elles-mêmes.**

²¹ Dans un tel contexte, un acteur ayant connaissance de l'existence d'une vulnérabilité n'aurait pas d'intérêt à conserver cette connaissance secrète car elle pourrait être exploitée par un rival. Cette co-construction permet d'établir une situation d'équilibre vertueux où chaque acteur a de très fortes incitations à rendre le dispositif de surveillance le plus fiable et le plus robuste possible.

²² Petrie et al. "Flexible Hardware-Enabled Guarantees for AI Compute." *ArXiv* (2025). <http://arxiv.org/abs/2506.15093>.

1. Solution immédiatement disponible : suivre l'activité des centres de calcul

Métrique d'intérêt : consommation électrique et température

La consommation électrique et la température dégagée sont deux métriques observables du niveau d'utilisation d'un centre de calcul, bien qu'elles ne donnent que très peu d'informations concernant les motifs de l'activité observée. En effet, il serait possible de moduler délibérément l'activité d'un cluster de calcul pour générer des motifs d'activité différents et tromper les auditeurs.²³

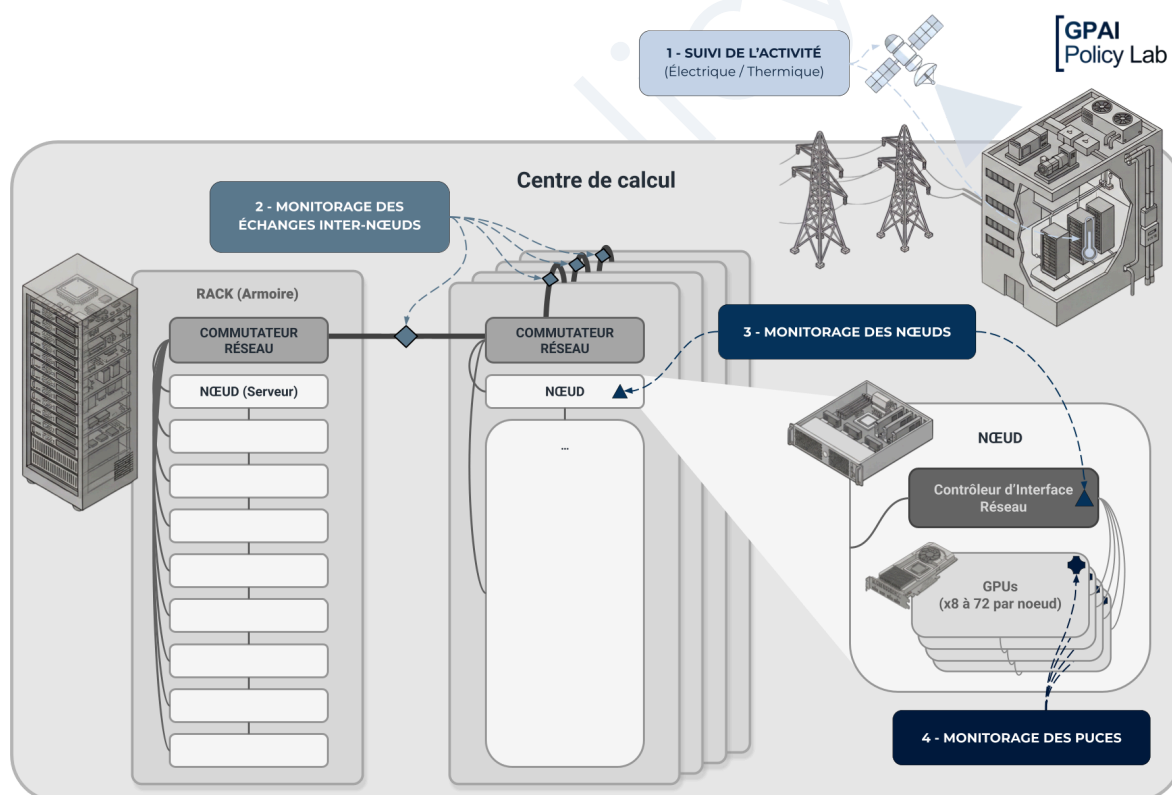
Malgré ces limitations fortes, disposer d'informations de ce type permettrait :

- de savoir si, et quand, un centre de calcul est utilisé;
- dans certains cas d'identifier des centres de calcul clandestins.

Ces informations seraient par ailleurs utiles à la mise en place de mécanismes de vérification plus avancés.

Dispositifs de vérification : mesures externes ou internes

Ces dispositifs de mesures peuvent prendre plusieurs formes : externes ou internes au centre de calcul (voir Figure 3).



gpaipolicylab.org

©

Figure 3 - Architecture d'un centre de calcul et insertion des solutions de vérification. Les bâtiments d'un centre de calcul contiennent des armoires (racks), dans lesquelles sont disposés les nœuds contenant les GPUs. Les GPUs d'un nœud transmettent leurs données via un contrôleur d'interface réseau, qui les échange ensuite via le commutateur réseau du rack. Les quatre approches de vérification envisagées se placent à ces différentes échelles de l'infrastructure physique du centre.

²³ De plus, il n'est pas possible de distinguer aisément de multiples calculs menés en parallèle d'un entraînement distribué exploitant l'intégralité des ressources d'un centre de calcul. Par exemple, du point de vue de la consommation électrique et de la température, 5 calculs parallèles exploitant chacun 20% du parc de GPUs d'un cluster ne sont pas distinguables d'un seul *training run* exploitant 100% du cluster.

Dans le cadre de dispositifs externes, on s'appuie sur les données mises à disposition par les gestionnaires de réseau ou sur des sondes électriques branchées sur l'alimentation des centres de calcul, et sur de l'imagerie satellite pour l'observation à la marge d'émission de température.

Dans le cadre de dispositifs internes, plus fiables, des sondes pourraient être branchées directement sur l'alimentation électrique de chaque *nœud* (serveur opérant entre 8 et 72 GPUs²⁴), et des sondes de température pourraient être installées auprès des systèmes de refroidissement (ou au sein des nœuds). **Cette option, plus ambitieuse, permettrait d'améliorer sensiblement la fiabilité des mesures d'activité, et peut déjà être déployée.**

2. Solution minimale : vérifier la puissance de calcul par des preuves cryptographiques

Métrique d'intérêt principale : nombre de FLOPs

La puissance de calcul utilisée pour entraîner les modèles d'IA est actuellement la meilleure métrique disponible pour prédire leurs performances tous domaines confondus. Cette puissance de calcul est quantifiée par le nombre de FLOPs (nombre d'opérations arithmétiques) utilisés au cours de l'entraînement d'un modèle. Le nombre de FLOPs a déjà été identifié comme métrique de référence au sein de plusieurs règlements et accords (AI Act²⁵, Executive Order 14110²⁶, California Bill SB 1047²⁷), ce qui en fait un bon candidat pour figurer au sein d'engagements internationaux.

Par ailleurs, d'autres propriétés des entraînements peuvent également être vérifiées, telles que le fait d'utiliser de l'apprentissage par renforcement ou bien certaines propriétés spécifiques des jeux de données utilisés (présence d'informations avancées accélérant le développement de capacités CBRN, informations protégées par les droits d'auteur).

Dispositif de vérification : observation réseau et preuves cryptographiques

Une approche pour l'instant sous-explorée consiste à s'appuyer sur des preuves cryptographiques dites "à divulgation nulle de connaissance"²⁸ (*zero-knowledge proofs*), pour permettre aux développeurs de modèles de déclarer leurs activités en envoyant des preuves issues de mesures dans leurs centres de calcul (cf. Encadré 5). Un tel type de preuve à divulgation nulle de connaissance aurait pu sembler inaccessible. **Pourtant il s'agit en réalité d'une technologie éprouvée** dont dépend le passage à l'échelle de nombreuses applications *blockchain*²⁹. Déployer cette approche en tant que mécanisme de vérification requiert quelques développements théoriques, mais surtout essentiellement applicatifs, pour réussir à intégrer les procédures d'entraînement (et en particulier les étapes de rétro-propagation) dans des preuves cryptographiques dont la complexité doit être maîtrisée. Les progrès à réaliser pour pouvoir mettre au point ce type de preuves impliquent une feuille de route estimée à 18 à 36 mois ; la durée des scénarios variant significativement en fonction de la méthode employée pour capturer les paquets échangés sur le réseau³⁰.

²⁴ On utilise ici GPU comme nom générique des puces de calcul IA, qui incluent d'autres architectures (TPU, NPU, etc.).

²⁵ "Regulation (EU) 2024/1689 ... (Artificial Intelligence Act). Official Journal of the European Union L 1689" *European Parliament and Council of the European Union*. (12 July 2024).

²⁶ "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Exec. Order No. 14110, 88 Fed. Reg. 75191" *Executive Office of the President of the United States*. (Nov. 1, 2023).

²⁷ www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

²⁸ "Senate Bill 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act. 2023–2024 Regular Session." *California Legislature*. leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047

²⁹ Lavin et al. "A Survey on the Applications of Zero-Knowledge Proofs." *ArXiv* (2024). arxiv.org/abs/2408.00243.

³⁰ Cette approche a par exemple déjà été intégrée à la blockchain Ethereum, notamment au niveau des *zk-rollup* et des machines virtuelles Ethereum à connaissance nulle (*zkEVM*) de couche 2 (L2).

³⁰ Une feuille de route technique détaillée est actuellement en préparation au GPAI Policy Lab.

Encadré 5 (détails techniques) - Preuves à divulgation nulle de connaissance appliquées à la vérification de l'entraînement de GPAI

Dans le cadre d'un entraînement distribué de modèle de frontière, un mécanisme de vérification intégrant des *zero-knowledge proofs* appliquerait la procédure suivante :

1. Les développeurs déclarent à l'avance leurs procédures de calcul dont la signature cryptographique est publique et dont le nombre de FLOPs est déclaré.
2. Ensuite, les paquets échangés entre les nœuds (informations circulant entre les GPU, cf. Figure 3) lors d'entraînements distribués peuvent être échantillonnés de manière aléatoire et leur signature cryptographique enregistrée.³¹
3. Les développeurs sont alors contraints de produire des preuves cryptographiques démontrant que les signatures observées ont bien été produites par la procédure déclarée initialement. Ce faisant, lesdites procédures et le reste des informations propriétaires ne sont jamais divulguées, excepté les signatures des paquets observés dont il n'est pas possible d'extraire des secrets industriels.

3. Solution avancée I : suivre les calculs au niveau des nœuds

Métrique d'intérêt principale : nombre de FLOPs

La métrique est la même que dans la solution précédente. Cette solution pourrait par ailleurs être plus aisément étendue à une gamme encore plus large d'autres métriques d'intérêt de manière flexible : les dispositifs sont prévus pour pouvoir vérifier de nouvelles propriétés à mesure que les méthodes d'entraînement et que la réglementation évoluent.

Dispositif de vérification :

Les dispositifs d'observation intra-nœud s'appuient essentiellement sur des modifications *hardware* des contrôleurs d'interface réseau (cf. Figure 3) afin d'y intégrer une micro carte de monitoring³². Les contrôleurs d'interface réseau orchestrent les communications entre les GPUs d'un nœud et le reste du cluster de calcul. De par leur positionnement sur les chemins de circulation des données, ces cartes disposent d'informations plus fines et seraient beaucoup plus robustes aux tentatives de modification par des acteurs malveillants³³.

Notamment, les contrôleurs d'interface réseau peuvent directement observer et collecter les métriques d'activité des GPUs de leur nœud. Cela permettrait à terme de se passer, au moins en partie, de systèmes de preuves cryptographiques pouvant être coûteux.

Une telle approche requiert le développement de solutions *hardware* dédiées, dépendantes de la coopération des fabricants actuels de ces composants de pointe. Elle nécessiterait, en raison du temps d'itération nécessaire à ces développements, une feuille de route estimée à 5 ans.

4. Solution avancée II : suivre les calculs au niveau des GPU

Métriques d'intérêt : nombre de FLOPs et traçabilité des opérations mathématiques

La métrique est la même que dans la solution précédente (FLOPs). S'ajoute à cela la possibilité de monitorer directement certaines opérations mathématiques au sein des GPUs. **Cela permettrait**

³¹ De multiples options sont disponibles pour produire ces sondes réseau : modification micrologicielle des contrôleurs d'interface réseau, modification des commutateurs réseau, conception de sondes physiques se branchant directement sur les câbles de communications.

³² Petrie, James, and Onni Aarne. "Technical Options for Flexible Hardware-Enabled Guarantees." *ArXiv*, (2025). arxiv.org/abs/2506.03409.

³³ La modification de composants soudés au sein des contrôleurs d'interface réseau est extrêmement délicate et risque d'endommager le contrôleur.

d'intégrer des propriétés spécifiques (procédures d'entraînement, architecture, etc.) au sein des engagements internationaux. Ainsi, les acteurs pourraient prouver qu'ils n'utilisent pas certaines méthodes d'optimisation ou certaines architectures dont les propriétés sont susceptibles de produire des IA présentant des comportements délétères identifiés dans les modèles de menace.

Dispositif de vérification :

Les approches proposant des dispositifs de monitoring intra-GPU se focalisent sur le fait d'intégrer une micropuce directement entre le circuit imprimé principal (PCB) et la mémoire (*high-bandwidth memory*) du GPU³⁴ (cf. Figure 3). Outre sa capacité de monitoring supérieure, l'avantage majeur de cette approche réside dans le fait qu'elle rend quasiment impossible toute tentative de modification par un acteur malveillant : le niveau de miniaturisation est tel que toute tentative pour retirer la micropuce risquerait d'endommager irréversiblement les GPUs équipés.

Il s'agit des scénarios les plus ambitieux nécessitant non seulement la coopération des fabricants de puces IA, mais également des efforts de R&D extrêmement conséquents pour intégrer de telles solutions sans générer aucune friction au niveau des échanges d'informations entre les cœurs du GPU et sa mémoire.

V. Un écosystème de vérification du compute est en train de se structurer, pour l'instant principalement au Royaume-Uni et aux États-Unis

Les solutions présentées dans les sections précédentes font l'objet d'un effort de R&D actif de la part d'un écosystème d'acteurs encore restreint. **Cet écosystème reste peu doté au regard des enjeux, mais il s'est structuré rapidement depuis 2023**, avec l'émergence d'organisations spécialisées, d'initiatives de financement dédiées et de premiers démonstrateurs. Il est actuellement concentré au Royaume-Uni et aux États-Unis, avec une présence européenne marginale.

Les acteurs les plus directement pertinents pour les solutions avancées décrites dans cette note s'organisent autour du concept de *Flexible Hardware-Enabled Guarantees (FlexHEGs)*, qui correspond aux approches intra-nœud et intra-GPU présentées plus haut. Conçus pour être résistants aux modifications par des acteurs disposant de moyens étatiques, ces composants s'intègrent aux équipements de calcul et permettent un monitoring transparent des entraînements. *Atlas Computing* (à but non lucratif, États-Unis) coordonne un programme de recherche associé.³⁵ *Ulyssean* (Royaume-Uni) développe une implémentation à l'échelle des centres de calcul, avec pour objectif affiché de couvrir l'ensemble de ceux des pays alliés des États-Unis d'ici 2028. *Amodo Design* (Royaume-Uni), fondée en 2023, assure l'ingénierie hardware et firmware de l'architecture *FlexHEG* et figure parmi les partenaires d'activation de l'*ARIA (Advanced Research and Invention Agency)*, aux côtés de Cambridge, Oxford et du MIT. D'autres approches se concentrent sur la couche physique de protection contre les altérations. *TamperSec* (Royaume-Uni, pré-seed) développe des prototypes capables de détecter les tentatives de modifications des systèmes de vérification installés dans les centres de calcul. *Earendil* (États-Unis) produit un système de destruction irréversible du hardware en cas d'accès non autorisé, conçu pour le renforcement des contrôles à l'export³⁶.

Pour les approches cryptographiques, *EZKL*³⁷ (États-Unis, 3,6 millions de dollars levés) est la seule librairie audité disponible pour la génération de preuves à divulgation nulle de connaissance

³⁴ Petrie, James. "Guaranteeable Memory: An HBM-Based Chiplet for Verifiable AI Workloads", *ICML Workshop on Technical AI Governance* (TAIG), 2024. openreview.net/forum?id=uc79kOv0MV

³⁵ "Announcing Flexible Hardware-Enabled Governors." *Atlas Computing* (2024). blog.atlascomputing.org/p/announcing-flexible-hardware-enabled.

³⁶ Earendil. earendil.ai.

³⁷ EZKL. Documentation. docs.ezkl.xyz.

appliquées aux modèles d'IA, relevant directement de la solution minimale décrite précédemment.³⁸ Un champ adjacent est celui du *confidential computing* (calcul en enclaves matérielles sécurisées), qui rassemble un nombre croissant d'acteurs principalement américains (*Tinfoil*, *Lunal*, *Lucid Computing*, *EQTY Lab* avec 16 millions de dollars en seed), auxquels s'ajoutent *Mithril Security* (France, rachetée en 2024 par H, laboratoire d'IA français) et *Edgeless Systems* (Allemagne). Ces acteurs adressent principalement des besoins de confidentialité des données en entreprise plutôt que la gouvernance internationale du compute au sens de cette note ; leurs solutions techniques restent néanmoins susceptibles de contribuer à certains dispositifs de vérification.

Sur le plan institutionnel, l'*Institute for AI Policy and Strategy* (IAPS, États-Unis) a analysé la possibilité d'utiliser des mécanismes de vérification de localisation des puces de calcul IA³⁹, et le *Center for a New American Security* (CNAS, États-Unis) a publié un rapport fondateur sur les puces gouvernables⁴⁰. La *RAND Corporation* a également contribué à la structuration du domaine en organisant en 2024 un atelier dédié à ces mécanismes et en publiant sur le sujet⁴¹.

Les principaux financeurs du domaine sont synthétisés dans le Tableau 1 ci-dessous.

Tableau 1 - Principaux financeurs de la vérification du compute (2024-2025).

ACTEUR	PAYS	MONTANT	NATURE	OBJET
<i>Advanced Research and Invention Agency (ARIA)</i>	Royaume-Uni	109 M £ (dont moins de 10% sur la vérification)	Agence publique	Programmes <i>Safeguarded AI</i> ⁴² et <i>Scaling Trust</i> ⁴³
<i>Institute for Progress (IFP)</i>	États-Unis	Projet à 30 M \$ ⁴⁴ ; jusqu'à 200 M \$ ⁴⁵	Think tank	Vérification par le hardware
<i>Survival and Flourishing Fund (SFF)</i>	États-Unis	4,1 M \$ (2024)	Philanthropique	Programme FlexHEGs ⁴⁶
<i>Coefficient Giving</i>	États-Unis	N.D.	Philanthropique	A financé RAND et le CNAS sur ces sujets ⁴⁷
<i>Longview Philanthropy</i>	Royaume-Uni	2-10 M \$ (2025)	Philanthropique	Appel à projets vérification par le hardware ⁴⁸

L'analyse de cet écosystème confirme une forte concentration autour du Royaume-Uni et des États-Unis, une contribution marginale de l'Union Européenne, et un niveau global de ressources encore très limité.

³⁸ La solution EZKL ne permet de prouver que l'inférence et non l'entraînement.

³⁹ Brass. "Location Verification for AI Chips." *IAPS* (2025). iaps.ai/research/location-verification-for-ai-chips.

⁴⁰ Aarne, Fist & Withers. "Secure, Governable Chips." *CNAS* (2024). cnas.org/publications/reports/secure-governable-chips.

⁴¹ Kulp, Siu, & Heim. "Insights from a Workshop on Hardware-Enabled Governance Mechanisms." *RAND Corporation* (2025). [rand.org/pubs/conf_proceedings/CFA3056-1.html](https://www.rand.org/pubs/conf_proceedings/CFA3056-1.html) ; Baker, Kulp, Marks, Brundage & Heim. "Verifying International Agreements on AI." *RAND Corporation* (2025). https://www.rand.org/pubs/working_papers/WRA4077-1.html.

⁴² "Safeguarded AI." *ARIA*. aria.org.uk/opportunity-spaces/mathematics-for-safe-ai/safeguarded-ai, voir notamment le rapport suivant : Petrie et al. "Flexible Hardware-Enabled Guarantees for AI Compute." *ArXiv* (2025). [http://arxiv.org/abs/2506.15093](https://arxiv.org/abs/2506.15093).

⁴³ "Scaling Trust." *ARIA*. aria.org.uk/opportunity-spaces/trust-everything-everywhere/scaling-trust.

⁴⁴ Ammann & Dalrymple. "Faster AI Diffusion Through Hardware-Based Verification." *Institute for Progress* (2025). ifp.org/faster-ai-diffusion-through-hardware-based-verification.

⁴⁵ Information communiquée au GPAI Policy Lab, non publiée à la date de rédaction.

⁴⁶ "SFF-2024 : Mechanisms for Flexible Hardware-Enabled Guarantees (FlexHEGs)." *Survival and Flourishing Fund* (2024). survivalandflourishing.fund/2024/flexhegs-recommendations.

⁴⁷ "AI Safety and Security Need More Funders." *Coefficient Giving* (2025). <https://coefficientgiving.org/research/ai-safety-and-security-need-more-funders>.

⁴⁸ "Request for Proposals on Hardware-Enabled Mechanisms (HEMs)." *Longview Philanthropy* (2025). <https://www.longview.org/request-for-proposals-on-hardware-enabled-mechanisms-hems-for-ai-verification>.

Conclusion

Le développement de systèmes d'IA avancés sans garanties de contrôle expose l'ensemble des acteurs internationaux à des conséquences irréversibles. La dynamique actuelle de course à l'AGI entre les États-Unis et la Chine aggrave cette exposition en dissuadant chaque acteur de prioriser unilatéralement la sécurité. La reconnaissance des enjeux de perte de contrôle favorise toutefois un basculement, depuis la dynamique de course, vers une dynamique de coordination. Une telle coordination permettrait de conditionner la poursuite du développement vers l'AGI à l'existence de garanties de contrôle solides.

Cette coordination ne peut reposer que sur des engagements assurés par des mécanismes de vérification crédibles. Ces mécanismes ne constituent pas une solution au problème fondamental de l'absence de garanties de contrôle des systèmes d'IA avancés, mais ils offrent des options réalistes pour que des acteurs rivaux collaborent en confiance sur ces enjeux.

Plusieurs solutions techniques permettant cette vérification sans divulguer d'informations confidentielles sont envisageables, selon des horizons de déploiement allant de l'immédiat à plusieurs années. Une solution minimale, fondée sur des preuves cryptographiques à divulgation nulle de connaissance, pourrait être opérationnelle sous 18 à 36 mois.

Dans ce contexte, la France dispose de plusieurs atouts pour mener cet effort. Le pays possède une expertise reconnue en cryptographie et en sécurité matérielle, un écosystème industriel en microélectronique (STMicroelectronics, CEA-Leti, Bull), un réseau de centres de calcul de taille croissante, une tradition d'engagement dans les régimes internationaux de non-prolifération et de vérification, ainsi qu'une position institutionnelle au sein de l'Union européenne lui permettant de porter ce type de sujets à l'échelle continentale. Si la France souhaite se positionner sur les mécanismes de vérification, cette situation lui offre l'option de s'affirmer comme pionnière en la matière.

Suite possible à cette note

1. Quelles sont les options opérationnelles dont la France dispose pour faire avancer le domaine des mécanismes de vérification ?
 2. Quels sont les prérequis nécessaires à la mise en place de mécanismes de vérification pour l'entraînement des GPAI de pointe à l'échelle internationale ?
 3. En quoi la structure de la chaîne de production des capacités de calcul conditionne-t-elle les leviers de coordination disponibles et les solutions possibles de mécanismes de vérification ?
-