General Purpose AI Policy Lab

Automatisation de la R&D en IA

Perspectives économiques, seuils critiques et leviers pour la France

Septembre 2026

#7 - Note

General Purpose AI Policy Lab

RÉSUMÉ EXÉCUTIF

— CONTEXTE —

Les systèmes d'IA de frontière participent déjà à la conception de la génération qui leur succède. L'automatisation complète de la R&D en IA n'est plus une perspective théorique mais un programme industriel en cours, soutenu par des investissements sans précédent. Si cette automatisation aboutit, elle rendrait toute anticipation de trajectoire à partir du rythme historique de progrès caduque, alors même que l'on ne dispose à ce jour d'aucune garantie de contrôle sur ces systèmes. Il existe cependant des leviers qui permettent d'infléchir la trajectoire actuelle, mais leur efficacité est contrainte par la rapidité à laquelle évoluent les capacités des systèmes d'IA à usage général (GPAI) de frontière.

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

— QUESTIONS —

- À partir de quels seuils de capacités les systèmes d'intelligence artificielle poseront-ils des problèmes de sécurité et de contrôle critiques ?
- À quel point l'automatisation de la R&D en IA actuellement en cours peut-elle accélérer le franchissement de ces seuils ?
- Quels sont les leviers qui permettraient de ne pas dépasser ces seuils avant d'avoir développé des garanties de contrôle ?

— SYNTHÈSE —

L'automatisation de la R&D en IA est un programme de recherche dont l'objectif est de concevoir des systèmes d'IA capables de réaliser en autonomie complète l'ensemble des tâches actuellement dévolues aux chercheurs et ingénieurs, c'est-à-dire de la conception même de systèmes d'IA. Dans l'hypothèse d'une automatisation totale, chaque génération de systèmes concevrait intégralement la génération suivante. Une telle boucle de rétroaction impliquerait une accélération substantielle du rythme de développement de systèmes plus performants, généraux et agentiques.

Ce programme est un objectif activement poursuivi par les entreprises qui conçoivent les IA de frontière. **Mi-2026, les systèmes d'IA de frontière participent déjà à la conception des générations suivantes** de systèmes plus avancés, ce qui produit une accélération mesurable du progrès des capacités des systèmes d'IA.

Une automatisation totale de la R&D est susceptible de produire une accélération majeure des capacités des systèmes d'IA. Il n'est pas possible d'anticiper correctement l'évolution des capacités futures à partir d'une simple extrapolation du rythme historique. Les capacités des systèmes d'IA les plus performants sont déjà suffisamment avancées pour présenter des menaces sur certains pans de la sécurité nationale (cybersécurité notamment). **Ne pas considérer l'automatisation de la R&D reviendrait à sous-estimer très largement les horizons temporels à partir desquels des seuils critiques de capacités seront franchis.**

L'automatisation de la R&D en IA va accélérer l'évolution des capacités des systèmes d'IA, avec pour conséquence la conception de systèmes plus agentiques, c'est-à-dire en capacité de poursuivre des objectifs complexes dans un environnement variable, sans supervision humaine et d'atteindre des niveaux de performance supérieurs aux experts dans l'ensemble des domaines cognitifs. Pourtant, si l'agentivité est une propriété recherchée parce qu'elle permet de réduire la supervision humaine, elle est également à l'origine d'un ensemble de risques amenés à se renforcer à mesure que des systèmes plus agentiques sont déployés.

Dans ce rapport, un système doté d'une agentivité suffisamment élevée pour être capable de performances supérieures aux experts dans tous les domaines cognitifs est nommé "AGI au seuil critique", ou AGI-SC. Des modélisations fondées sur la théorie économique de la croissance sont utilisées pour anticiper les effets d'une automatisation de la R&D en IA sur l'accélération des capacités et le franchissement de ce seuil. **L'analyse de trois de ces modélisations (GATE, AI Futures, Software Intelligence Explosion) indique que sans actions limitantes, la moitié des trajectoires simulées par ces trois modélisations franchissent le seuil critique (AGI-SC) avant fin 2034 et environ un quart avant 2030.**

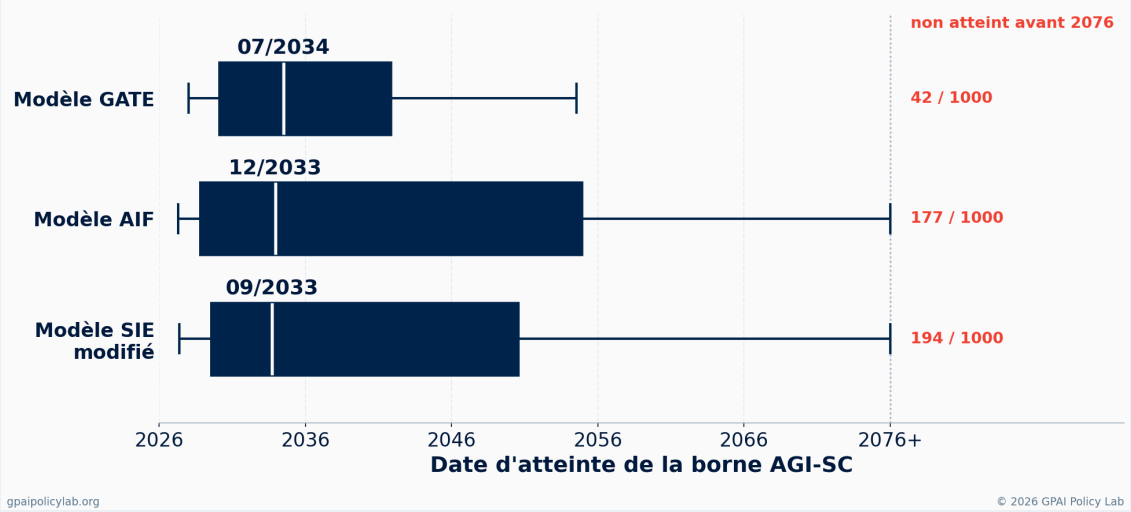

Figure A - Résultats d'harmonisation des trois modélisations analysées dans ce rapport. La date médiane de franchissement du seuil AGI-SC est représentée pour chacune sur un ensemble de 1000 trajectoires, avec l'intervalle de confiance à 80%.

General
Purpose AI
Policy Lab

— IMPLICATIONS POUR LES ACTEURS FRANÇAIS ET EUROPÉENS —

Il n'existe à l'heure actuelle aucun mécanisme de gouvernance permettant de faire face à des transformations technologiques d'ampleur dans un temps aussi restreint, rendant incertaine l'adaptation des institutions et des sociétés.

L'automatisation de la R&D est déjà en cours, avec une forte opacité et une absence de supervision

L'absence de transparence des entreprises à la frontière sur l'état actuel de l'automatisation de la R&D en IA place la France et l'Europe dans une situation d'asymétrie d'information, limitant leur capacité à anticiper et à prévenir les atteintes à la sécurité nationale et internationale.

Les entreprises procèdent activement à de la rétention d'informations concernant les capacités de leurs systèmes d'IA en R&D. En l'absence d'accès direct à ces données, on peut s'attendre à ce que les modélisations sous-estiment la quantité de trajectoires rapides (AGI-SC franchi avant 2030), qui constituent déjà un quart de toutes les simulations.

La France ne dispose d'aucun accès aux informations relatives au développement des IA de frontière et à leurs capacités dans des domaines susceptibles de porter atteinte à sa sécurité nationale. Ce déficit d'information serait inconcevable au regard des standards en vigueur dans d'autres industries à risque comme l'aviation, le médical ou le nucléaire civil. **Des exigences de transparence constituent donc un prérequis pour que la France puisse appréhender au mieux les conséquences imputables au développement de ces systèmes en dehors de son territoire national.**

Par ailleurs, les modélisations dans lesquelles la puissance de calcul est plafonnée (pour l'entraînement des modèles et la R&D en IA) éloignent la date de franchissement du seuil critique (AGI-SC). Plus cette limitation intervient tôt, plus ses chances de succès sont élevées. À l'heure actuelle, 86% de la puissance de calcul mondiale est possédée par les États-Unis, contre moins d'1% pour la France, impliquant qu'actionner ce levier nécessite prioritairement des mesures coordonnées pour peser en dehors du territoire national.

General
Purpose AI
Policy Lab

Scénarios d'automatisation de la R&D en IA

Évolution des capacités, risques émergents et leviers pour la France

Ce rapport vise à répondre à 3 questions :

- À partir de quels seuils de capacités les systèmes d'intelligence artificielle poseront-ils des problèmes de sécurité et de contrôle critiques ?
- À quel point l'automatisation en cours de la R&D en IA est-elle susceptible d'accélérer l'atteinte de ces seuils de capacités ?
- Quels sont les leviers qui permettraient de ne pas dépasser ces seuils avant d'avoir développé des garanties de contrôle ?

L'automatisation d'une tâche correspond à la délégation à une machine du travail auparavant accompli par des humains. Plusieurs précédents d'automatisation ont marqué l'histoire industrielle, en se concentrant d'abord sur des tâches physiques, permettant de produire plus vite, à moindre coût et avec une plus grande efficacité. Dans la lignée des premiers ordinateurs, l'apparition de systèmes d'intelligence artificielle à usage général (*General Purpose Artificial Intelligence*, ou GPAI) étend le champ de l'automatisation à une large gamme de tâches cognitives. Certaines tâches y échappent encore, mais la liste se réduit à chaque nouvelle génération de modèles.

L'un des domaines dont l'automatisation provoquerait des transformations majeures est la recherche et développement (R&D) en IA, c'est-à-dire de la conception même de systèmes d'IA. Dans un contexte de compétition intense entre les entreprises à la frontière (OpenAI, Anthropic, Google DeepMind, etc.), sur fond de rivalité technologique internationale, **l'automatisation de la R&D en IA est présentée par plusieurs acteurs de premier plan comme une source d'avantage stratégique potentiellement décisif.** Dans l'hypothèse d'une automatisation totale, chaque génération de systèmes concevrait intégralement la génération suivante. Une telle boucle de rétroaction impliquerait une accélération substantielle du rythme de développement de systèmes plus performants, généraux et agentiques[1]. Le premier acteur à enclencher une telle dynamique pourrait prendre une avance difficile, voire impossible, à rattraper par ses concurrents - d'où l'enjeu stratégique qu'elle représente.

L'automatisation de la R&D en IA constitue désormais un programme de recherche activement poursuivi par les entreprises développant les systèmes les plus avancés. Quantifier précisément l'ampleur de cette accélération, les enjeux et risques associés, ainsi que l'efficacité de diverses

[1] L'agentivité est ici définie comme la capacité d'un système à atteindre un objectif dans un environnement complexe et perturbé, avec une supervision minimale et une capacité de planification et de gestion de l'imprévu. Il existe un gradient d'agentivité, et les systèmes actuels sont relativement peu agentiques. D'après Chan et al. "Harms from Increasingly Agentic Algorithmic Systems." FAccT '23, pp. 651–666 (2023). https://doi.org/10.1145/3593013.3594033.

mesures de prévention potentielles est essentiel, puisque l'arrivée de systèmes capables d'égaler ou de surpasser les meilleurs experts sur l'ensemble des tâches cognitives n'affecte pas les sociétés avec la même intensité selon qu'elle se produit en quelques années ou en plusieurs décennies.

Le présent rapport tente de répondre à ces questions en analysant et comparant trois modélisations fondées sur la littérature économique, pour anticiper la rapidité des progrès des systèmes d'IA sous l'hypothèse d'une automatisation partielle puis totale de la R&D. L'évolution des capacités est mise en regard des risques que présentent des systèmes d'IA dépassant certains seuils de performance et d'autonomie.

I. À partir de quels seuils de capacités les systèmes d'intelligence artificielle poseront-ils des problèmes de sécurité et de contrôle critiques ?

Depuis le premier déploiement public de ChatGPT en novembre 2022, les capacités des systèmes d'intelligence artificielle se sont améliorées à un rythme rapide, tant en performance qu'en généralité[2]. Au cours des deux dernières années, l'arrivée de modèles dits « de raisonnement » a encore accru cette progression[3], et les modèles les plus performants à la mi-2026 sont déjà capables de surpasser les performances des experts humains sur certaines tâches de certains domaines, comme la programmation informatique compétitive[4] ou les mathématiques[5]. La rapidité à laquelle les capacités progressent soulève la question des enjeux de sécurité qui pourraient être amenés à émerger ou à se renforcer à mesure que les meilleurs systèmes d'IA deviennent plus performants, généraux et agentiques[6]. **Les seuils exacts de capacités à partir desquels émergeront des enjeux majeurs pour la sécurité nationale et internationale sont actuellement inconnus, ce qui rend difficile leur anticipation.**

1. Une harmonisation des définitions s'impose pour lier seuils de capacités et niveaux de risque

A. Les capacités des systèmes d'IA progressent rapidement et surpassent déjà celles des meilleurs experts dans certains domaines

Une manière de mesurer les capacités des systèmes d'intelligence artificielle est de concevoir des tests standardisés (*benchmarks*) qui évaluent des compétences données (raisonnement, connaissances générales ou spécialisées, etc.). Ces *benchmarks* permettent ensuite d'observer l'évolution des performances atteintes par les générations d'IA successives. On constate aujourd'hui un phénomène généralisé de saturation des benchmarks : les GPAI les plus avancées atteignent des

[2] Epoch AI. "eci-public: Epoch Capabilities Index." GitHub (2025). https://github.com/epoch-research/eci-public
[3] Ho et Berg. "Quantifying the algorithmic improvement from reasoning models." Epoch AI, Gradient Updates (2025). https://epochai.substack.com/p/quantifying-the-algorithmic-improvement
[4] Anshad. "OpenAI's Reasoning Model Swept AtCoder 2026, Beating Every Human Competitor." terminalblog (2026). https://terminalblog.com/blog/openai-atcoder-2026-ai-beats-humans/
[5] Epoch AI, « FrontierMath: Open Problems », benchmark de problèmes ouverts de mathématiques de recherche, 50 problèmes en juillet 2026 dont 5 résolus par IA https://epoch.ai/frontiermath/open-problems
[6] Chan et al. "Harms from Increasingly Agentic Algorithmic Systems." FAccT '23, pp. 651–666 (2023). https://doi.org/10.1145/3593013.3594033

scores proches du maximum théorique sur la majorité des tests, y compris ceux conçus spécifiquement pour être difficiles à résoudre par des systèmes d'intelligence artificielle[7].

Ces progrès rapides se traduisent notamment par une amélioration des capacités agentiques, c'est-à-dire la faculté de compléter des tâches longues, nouvelles et complexes sans supervision humaine. L'une des métriques utilisées pour mesurer ces capacités est « l'horizon temporel » : **il s'agit de mesurer le temps de complétion par des experts en informatique d'une série de tâches de complexité croissante, puis de regarder parmi cet ensemble de tâches quelles sont celles qui peuvent être complétées de manière autonome par des systèmes d'IA**. L'intérêt de cette métrique réside dans son caractère réaliste, ces tâches nécessitant un certain degré de planification et d'exploration. L'évolution des performances des GPAI sur cette métrique a suivi une croissance exponentielle sur les 7 dernières années.

Tableau 1 - Aperçu de l'évolution des capacités des modèles d'IA. L'horizon temporel est mesuré par rapport au temps de complétion de différentes tâches par des experts humains, sur lesquelles le système d'IA est ensuite évalué. On calcule ensuite la durée qui correspond à un taux de succès de 50%[8]

MODÈLE	DATE DE SORTIE	HORIZON TEMPOREL	CAPACITÉS OBSERVÉES
GPT-2	Février 2019	~2s	Peut produire quelques phrases cohérentes à la suite. Multiplie les contradictions factuelles et les erreurs basiques.
GPT-3	Juin 2020	~10s	Rédige des paragraphes cohérents. Traduit approximativement. Peut répondre à des questions simples mais échoue aux raisonnements en plusieurs étapes.
GPT-4	Mars 2023	~5 min	Réussit l'examen du barreau américain dans le top 25%. Écrit quelques lignes de code fonctionnelles. Résout des problèmes mathématiques de niveau collège. Fait preuve de capacités émergentes de raisonnement basiques.
GPT-5	Août 2025	~2h	Résout de manière autonome certaines tâches d'ingénierie logicielle qui prennent 2 à 3 heures à un développeur expérimenté. Obtient un score de 25% sur *Humanity's Last Exam* (benchmark pluridisciplinaire conçu pour être extrêmement difficile).
Claude Opus 4.5	Novembre 2025	~5h	Exécute des tâches d'ingénierie logicielle complexes. Anthropic indique que le modèle est utilisé en interne pour accélérer la recherche en IA elle-même.
Claude Mythos	Avril 2026	>16h	Découvre de manière autonome des vulnérabilités dans la majorité des grands systèmes d'exploitation et navigateurs. Anthropic a choisi de ne pas le diffuser publiquement en raison des risques cyber majeurs présentés.

On observe également des systèmes spécialisés capables d'un niveau de performance inatteignable pour des humains, souvent dans des domaines régis par un ensemble de règles bien définies (échecs, poker, jeu de Go). Pour autant, les GPAI de frontière actuelles atteignent des niveaux proches des meilleurs spécialistes dans des domaines plus complexes comme la

[7] "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI" ; Andréoletti et al. "Position: The Window for Preventive Action Before AI Saturates Most Cognitive Benchmarks is Closing" HAL (2026). https://hal.science/hal-05500135.

[8] Kwa et al. 'Measuring AI Ability to Complete Long Software Tasks." *NeurIPS* (2025). https://proceedings.neurips.cc/paper_files/paper/2025/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html., données disponibles : https://metr.org/time-horizons/

programmation, la résolution de problèmes mathématiques et la manipulation du langage[9]. Rien n'indique que les meilleures performances humaines soient optimales ou proches des limites atteignables ; si ce n'est pas le cas, cela signifie qu'un système d'IA dont la performance égale celle des meilleurs experts dispose encore d'une marge importante d'amélioration.

Encadré 1 - Performances surhumaines

La notion de performance surhumaine est déjà une réalité dans certains domaines. Aux échecs, le niveau des meilleurs systèmes d'IA est tel qu'aucun professionnel ne peut espérer gagner contre un ordinateur. On a pu observer temporairement l'émergence du « *centaur chess* » : un joueur humain s'appuie sur un programme d'échecs pour guider ses décisions, afin que la combinaison de l'intuition humaine et de la puissance de calcul des machines produise un niveau de jeu supérieur. Cependant, dès le début des années 2010, les programmes sont devenus si performants que l'intervention humaine a cessé d'apporter un avantage. Le passage par une phase de complémentarité homme-machine peut donc se révéler transitoire.

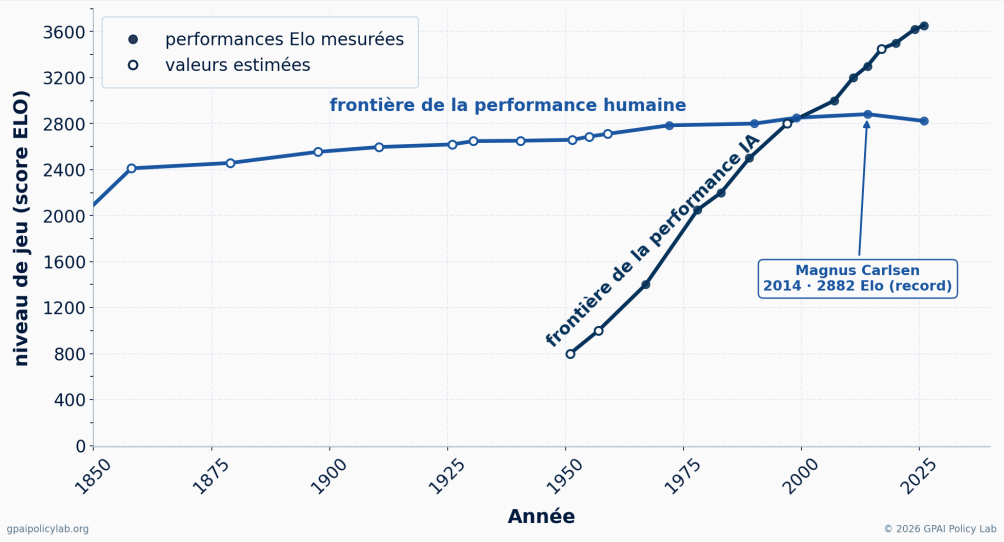

Figure 1 - Évolution des performances humaines et IA aux échecs, en ELO. Les performances historiques sont estimées par réanalyse de parties[10].

Aujourd'hui les niveaux surhumains de performance ne s'observent plus uniquement dans le domaine des jeux. En 2021, AlphaFold a prédit la structure tridimensionnelle de plusieurs millions de protéines, résolvant un problème fondamental de la biologie sur lequel des chercheurs travaillaient depuis cinquante ans[11]. En mai 2026, GPT-5.5 Pro a résolu une conjecture mathématique ouverte depuis 1946, en trouvant une solution dont

[9] GPAI Policy Lab. "Situer les modèles d'IA sur l'échelle de l'expertise humaine." Mars 2026. https://gpaipolicylab.org/blog-1.

[10] La comparaison entre les valeurs extrêmes d'ELO (>3000) atteintes par les meilleurs systèmes d'IA actuels et les scores humains est difficile (pas d'échelle commune), mais l'on a la certitude depuis 2006 que les meilleurs joueurs humains ne peuvent gagner aucun match contre les meilleurs systèmes d'IA (match Kramnik - Deep Fritz, perdu 4-2). Source des données disponible sur Github.

[11] Jumper et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

l'originalité a été reconnue par les spécialistes[12]. Dans ces deux cas, un système d'intelligence artificielle a résolu des problèmes qui avaient tenu en échec les meilleurs experts humains. La notion de performance surhumaine est donc déjà un fait établi dans des domaines aux applications réelles.

Les progrès des capacités sont rapides, mais il n'existe pas à l'heure actuelle d'échelle formalisée qui permette d'évaluer les risques associés à un niveau donné de performance, généralité et d'agentivité pour un système d'IA. Le terme "AGI" (pour *Artificial General Intelligence*, ou intelligence artificielle générale) est communément employé pour désigner des systèmes dont les capacités dépassent un seuil considéré comme significatif, mais l'usage de ce terme reste imprécis et implique une dichotomie qui manque de granularité.

B. Les définitions employées pour qualifier les capacités des systèmes d'IA sont imprécises.

Malgré son usage répandu, le terme d'AGI ne fait l'objet d'aucune définition consensuelle et recouvre plusieurs critères parfois très différents. L'usage de l'adjectif « général » est utilisé en opposition au caractère hautement spécialisé des premiers systèmes d'IA, capables d'exceller uniquement dans leur domaine spécifique (échecs, go, reconnaissance d'images). À l'inverse, une intelligence artificielle générale serait en mesure de réaliser un large éventail de tâches cognitives diverses. La polysémie du terme conduit à des évaluations divergentes sur la temporalité de développement d'une potentielle AGI et donc sur les enjeux de sécurité à anticiper dans un futur proche. Il n'existe présentement aucune échelle formelle communément admise qui permette d'évaluer le degré de généralité d'un système. S'ajoute à cela un glissement progressif des critères requis pour atteindre le seuil d'AGI à mesure que les systèmes d'IA franchissent des seuils de performance autrefois considérés comme caractéristiques d'une intelligence générale. Ainsi, la victoire aux échecs, puis au jeu de Go, puis la maîtrise du langage naturel, la réussite à des examens professionnels ou universitaires de haut niveau ont été successivement présentées comme des marqueurs décisifs de l'intelligence artificielle générale avant d'être requalifiées de simples prouesses techniques une fois franchies. Ce déplacement continu des critères rend d'autant plus difficile l'établissement de repères stables.

Encadré 2 - De multiples définitions du terme AGI

Le terme AGI diffère en signification selon les auteurs et les contextes. On peut distinguer plusieurs types de définitions, qui ne se recoupent que partiellement.

- **Définition psychométrique.** La notion d'AGI peut aussi être entendue comme un système d'IA d'une polyvalence cognitive équivalente à celle d'un adulte instruit, mesurée sur la base de critères psychométriques (intelligence fluide, mémoire long-terme, théorie de l'esprit, raisonnement mathématique, etc.). Appliquée aux

[12] Alon et al. "Remarks on the disproof of the unit distance conjecture." arXiv (2026). https://doi.org/10.48550/arXiv.2605.20695

modèles récents, cette grille fait apparaître un profil « en dents de scie », performant sur les tests de connaissance mais encore limité sur des fonctions comme la mémoire à long terme[13].

- **Définition économique.** *OpenAI* définit l'AGI comme un système « hautement autonome surpassant les humains dans la plupart des travaux à valeur économique[14] ». Le critère porte ici sur la capacité de systèmes à se substituer de manière rentable au travail humain. L'ambiguïté réside dans l'écart qui peut exister entre l'existence de systèmes capables d'automatiser une grande partie des tâches économiques et leur déploiement effectif, probablement ralenti par des contraintes externes.
- **Définition graduée et multidimensionnelle.** Les chercheurs de *Google DeepMind*[15] écartent l'idée d'un seuil binaire et proposent une matrice croisant le degré de généralité (étroit - pour un système spécialisé dans une seule tâche -, ou général - c'est à dire performant dans un large ensemble de domaines) avec cinq niveaux de performance (émergent, compétent, expert, virtuose, surhumain), calibrés par rapport à une population d'adultes éduqués.

Ces définitions divergent sur la nature même du critère retenu (cognition générale, autonomie, valeur économique, profil psychométrique), ce qui explique qu'un même système puisse être qualifié d'AGI selon l'une de ces définitions tout en demeurant éloigné de ce seuil selon une autre.

Les systèmes d'IA actuellement utilisés massivement par le grand public sont qualifiés de GPAI (*General Purpose AI*, IA à usage général) par opposition aux systèmes spécialisés, en raison de leur large domaine d'application (rédaction de récits fictifs, écriture de code informatique, capacités avancées en sciences et en droit, etc.). Malgré ce haut degré de généralité, et des performances parfois supérieures aux experts sur certaines tâches spécifiques en mathématiques ou en programmation, ils ne sont pas considérés comme atteignant le seuil AGI, du fait de limites importantes sur les tâches qui demandent une planification à long terme dans des environnements fortement variables (interactions nombreuses et complexes avec d'autres agents, nécessité d'explorer plusieurs options et de prendre des décisions sous forte incertitude, etc.)[16].

C. L'agentivité des systèmes d'IA est une caractéristique déterminante pour établir des seuils de risque

Évaluer l'étendue des capacités de systèmes d'IA s'avère en pratique ardu[17], car de nombreux facteurs amènent à sous-estimer les performances maximales du modèle (on parle d'*elicitation gap*). Par exemple, les systèmes les plus avancés sont partiellement limités par leur budget d'inférence,

[13] Hendrycks et al. "A Definition of AGI." arXiv (2025). https://arxiv.org/abs/2510.18212 ; Zhou et al. "General Scales Unlock AI Evaluation with Explanatory and Predictive Power." arXiv (2025). https://doi.org/10.48550/arXiv.2503.06378
[14] OpenAI. "OpenAI Charter." OpenAI (2018). https://openai.com/charter.
[15] Morris et al. "Levels of AGI for Operationalizing Progress on the Path to AGI." arXiv (2023). https://arxiv.org/abs/2311.02462
[16] Ou & Burnham. "AI doesn't get better at this board game with practice." Epoch AI (2026). https://epoch.ai/publications/earthborne-rangers-benchmark
[17] "Évaluation des IA de frontière : Des difficultés croissantes" GPAI Policy Lab (2026).

c'est-à-dire que les performances mesurées sont contraintes par la quantité de calcul allouée au système pendant l'évaluation[18]. Par ailleurs, un système peut présenter des niveaux de performances très élevés dans un domaine sur une tâche courte, mais se dégrader nettement lorsque les tâches nécessitent une planification à long terme[19]. Un système pleinement général doit être en mesure d'acquérir des connaissances nouvelles et d'interagir avec des environnements complexes, tout en maintenant une capacité d'adaptation élevée.

Ces caractéristiques peuvent être décrites par le terme "agentivité", défini plus généralement comme la capacité d'un système (avec ou sans architecture de support, ou *scaffolding*) à atteindre un objectif dans un environnement complexe[20]. Ainsi, les systèmes hautement spécialisés (conçus pour un unique type de tâches) sont peu agentiques, incapables d'acquérir et de transférer des connaissances dans un large ensemble de domaines, ce qui constitue un prérequis pour opérer sans aucune supervision humaine[21]. Pour évaluer le niveau d'agentivité d'un système, on peut étudier sa capacité à réaliser des tâches longues, nouvelles et complexes, sans supervision humaine. L'échelle qui en résulte propose 6 niveaux regroupés dans les trois classes suivantes :

- **Les systèmes faiblement agentiques** (A1, A2), capables d'interactions limitées et de performances moyennes dans des domaines cognitifs à faible complexité.
- **Les systèmes agentiques limités** (A3, A4), capables de planifier et d'exécuter des tâches en environnement connu, avec un excellent niveau de performance mais manquant de capacités d'exploration et de cohérence à long terme.
- **Les systèmes AGI au seuil critique** (A5 et A6), capables d'opérer sans aucune supervision humaine sur des tâches longues et complexes, à des niveaux de performance équivalents ou supérieurs aux experts humains.

Tableau 2 - Échelle d'agentivité et de risques. Tableau synthétisant les différents niveaux d'agentivité et les capacités qui y sont associées. [Voir annexe E pour plus de détails sur la construction de l'échelle]

NIVEAU DE L'ÉCHELLE	DESCRIPTION DU NIVEAU D'AGENTIVITÉ	DOMAINES COGNITIFS / TÂCHES RÉALISABLES	EXEMPLES
Systèmes non agentiques			
A0. IA-Outil	Pas d'autonomie. Le système génère une sortie (texte, image, classification) sans action externe ni interaction avec d'autres acteurs.	Jeux aux règles bien définies, classification d'image / de texte.	Google Traduction (Neural Machine Translation, 2016) DALL-E (Open AI, 2021)
Systèmes faiblement agentiques			
A1. Système conversationnel	Le système engage un dialogue cohérent avec l'humain, maintient le contexte d'une conversation et peut demander des précisions, mais n'agit pas au-delà de l'échange verbal.	Traduction simple, rédaction de textes non techniques, interactions simples avec un utilisateur.	Première génération de LLM (ChatGPT, 2022)

[18] AI Security Institute. "Evidence for inference scaling in AI cyber tasks: Increased evaluation budgets reveal higher success rates." Blog AISI (2026). https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates
[19] Rabanser et al. "Towards a Science of AI Agent Reliability." arXiv (2026) https://arxiv.org/abs/2602.16666.
[20] Chan et al. "Harms from Increasingly Agentic Algorithmic Systems." FAccT '23, pp. 651–666 (2023). https://doi.org/10.1145/3593013.3594033
[21] Chan et al. " op. cit.

A2. Agent minimal	Le système est capable de planifier sur des temps courts, peut enchaîner plusieurs actions pour accomplir une tâche bien définie et dans un environnement simple, en s'appuyant sur des outils externes (web, code, API).	Code informatique simple, Raisonnement et mathématiques de niveau lycée, aide au diagnostic médical à partir de symptômes, aide à la recherche d'informations.	LLMs avec chaîne de raisonnement (o1, dec. 2024 ; DeepSeek R1, 2025)
Systèmes agentiques limités			
A3. Agent planificateur	Le système peut accomplir une mission bien définie qui pourrait prendre plusieurs heures à jours pour un spécialiste humain. La fiabilité sur les tâches mal spécifiées, non vérifiables et/ou nécessitant un comportement exploratoire est encore très faible.	Code informatique complexe, tâches de secrétariat simples, mathématiques et science de niveau universitaire, droit, négociation commerciale en environnement simple.	LLMs avancés avec scaffolding (Opus 4.5 et au-delà, GPT5.2 et au-delà)
A4. Agent orchestrateur	Le système prend en charge des missions complexes pouvant prendre plusieurs jours à semaines à des équipes humaines. Il coordonne d'autres systèmes et des outils multiples, peut amorcer des comportements exploratoires et ajuster ses stratégies face aux obstacles. Les tâches longues en environnements très complexes sont encore une limite.	Recherche approfondie de vulnérabilités cyber, projets d'ingénierie informatique, aide majeure à la R&D dans certains domaines, droit et négociation commerciale en environnements complexes. Niveau proche d'experts humains dans certains domaines.	Incertain. Les modèles les plus avancés (Mythos, GPT-6) présentent peut-être des caractéristiques à la frontière de cette classe.
Systèmes AGI au seuil critique (AGI-SC)			
A5. Agent autonome (AGI)	Le système conduit des missions pouvant prendre plusieurs semaines à mois à des équipes humaines, à partir d'objectifs stratégiques. Il redéfinit ses sous-objectifs selon les circonstances, et opère dans des environnements variés comme un acteur autonome, en interagissant avec d'autres agents, individus et institutions.	Niveau équivalent voire supérieur à celui d'experts humains dans la majorité des domaines cognitifs, y compris sur des tâches longues et complexes.	Non atteint à ce stade
A6. Agent autonome surhumain (AGI Forte)	Le système est en mesure d'atteindre des objectifs stratégiques pouvant prendre des années à des organisations humaines, dans tous types d'environnements (y compris fortement dégradés et/ou nouveaux) en coordonnant/interagissant un grand nombre d'autres agents ou acteurs humains / institutionnels.	Niveau surhumain dans l'intégralité des domaines cognitifs.	Non atteint à ce stade

Dans ce rapport, on utilisera la terminologie « AGI » en ajoutant le qualificatif « au seuil critique » (soit AGI-SC) pour désigner les systèmes à partir de la catégorie "agent autonome" (A5). Un système relevant de cette catégorie dispose d'une agentivité suffisante pour atteindre des niveaux de performance équivalents voire supérieurs à ceux des meilleurs experts dans tous les domaines cognitifs, y compris sur des tâches très longues et complexes.

2. Le franchissement des seuils de capacités intensifie des enjeux déjà apparents

Les systèmes AGI au seuil critique (AGI-SC) sont caractérisés par un niveau d'agentivité et de performance susceptible d'intensifier fortement les enjeux qui émergent déjà avec les GPAI actuelles. Trois catégories de risques voient leur probabilité d'occurrence et leur sévérité s'accroître à mesure que les capacités des GPAI augmentent, et ces risques peuvent se matérialiser bien avant l'atteinte de la catégorie AGI-SC.

A. Perturbations sociales et économiques rapides

La capacité des GPAI à réaliser de manière autonome des tâches de plus en plus longues signifie qu'une part croissante des missions réalisées par des travailleurs humains est amenée à être transférée vers des systèmes d'IA, sous l'effet des pressions concurrentielles et des logiques de réduction de coût. Si les progrès sont extrêmement rapides, la transition ne pourra être absorbée par les mécanismes classiques de reconversion professionnelle[22]. Alors que les révolutions industrielles passées se sont étalées sur des décennies, laissant le temps aux systèmes éducatifs et aux marchés du travail de s'ajuster, un progrès concentré sur quelques années affecterait une proportion importante du secteur tertiaire, le travail cognitif étant le plus exposé à l'automatisation.

Cette dynamique pourrait s'accompagner d'une concentration de la valeur économique sans précédent, car le développement des systèmes d'IA de frontière requiert des capitaux et des infrastructures que seules quelques entreprises dans le monde (essentiellement américaines et dans une moindre mesure, chinoises) sont en mesure de mobiliser[23]. Si ces mêmes entreprises captent une part significative de la valeur produite par l'automatisation du travail, la concentration de richesses qui en résulterait dépasserait largement tout niveau historiquement observé[24]. Par ailleurs, l'asymétrie entre les pays qui possèdent les systèmes d'IA les plus avancés et ceux qui en sont dépourvus générerait des dynamiques d'extrême dépendance stratégique : être en possession de ces systèmes permettrait un développement technologique global fortement accéléré, y compris dans le domaine de la défense, accentuant les écarts de puissance déjà existants - potentiellement de façon irréversible. **Cette dynamique conduirait à une vassalisation, et pourrait rendre la notion même de souveraineté nationale caduque en raison du différentiel de capacités militaires, scientifiques et économiques.**

B. Mésusages par des acteurs malveillants

La performance des GPAI de frontière dans le domaine de la cybersécurité a franchi un seuil en 2026, illustré par les performances de Claude Mythos[25]. Anthropic a pris la décision inhabituelle de ne pas déployer publiquement ce système d'IA, estimant qu'il franchissait le seuil à partir duquel ses capacités offensives dépassent les capacités défensives disponibles dans l'écosystème. **Rien n'indique que ce niveau constitue un plafond, et la trajectoire observée suggère à l'inverse que les générations suivantes de modèles disposeront de capacités cyber**

[22] Korinek et Suh. "Scenarios for the Transition to AGI." arXiv (2024). https://doi.org/10.48550/arXiv.2403.12107

[23] 70% de croissance annuelle des dépenses dans les centres de données. Juniewicz. "Hyperscaler capex has quadrupled since GPT-4's release." Epoch AI, Data Insights (2026). https://epoch.ai/data-insights/hyperscaler-capex-trend

[24] Sytsma. "Decisive Economic Advantage: Modeling the Transition from Temporary First-Mover Leads to Economic Dominance in Artificial General Intelligence." RAND Corporation (2026). https://www.rand.org/pubs/research_reports/RRA4444-1.html

[25] Chauvin et al. "Are Mythos' cyber capabilities overhyped?" Epoch AI, Gradient Updates (2026). https://epoch.ai/gradient-updates/are-mythos-cyber-capabilities-overhyped

nettement supérieures[26]. Ces capacités devraient par ailleurs être répliquées rapidement (probablement courant 2027) par des GPAI *open-source*[27]. De telles capacités servent deux catégories d'acteurs : d'une part, les acteurs étatiques déjà dominants, qui disposeraient alors d'un multiplicateur de puissance, renforçant encore leur supériorité stratégique ; d'autre part, des acteurs infra-étatiques (groupes terroristes, organisations criminelles transnationales, acteurs isolés) qui disposaient jusqu'ici d'une capacité de nuire limitée, mais pourront désormais accéder, via les modèles *open source*, à des vecteurs d'attaque auparavant réservés aux grandes puissances.

Dans le domaine de la biosécurité, les capacités des GPAI progressent tout aussi rapidement, dépassant déjà le niveau d'experts sur de nombreux benchmarks[28] et il est envisageable que des modèles plus avancés puissent permettre à des attaquants peu expérimentés de produire des agents pathogènes existants et à des acteurs étatiques malveillants de concevoir des pathogènes nouveaux et particulièrement létaux[29].

C. Perte de contrôle liée aux comportements non alignés des modèles

Le terme « aligné » est utilisé pour qualifier un système d'IA qui poursuit effectivement les objectifs souhaités par ses concepteurs ou utilisateurs. Le paradigme actuel d'entraînement des IA frontière ne permet pas de produire des systèmes alignés : il ne permet pas de spécifier directement les objectifs d'un système ni de s'assurer qu'il les a effectivement intériorisés[30]. Le phénomène de mésalignement est d'ores et déjà documenté empiriquement[31] sur les systèmes actuels et constitue un enjeu scientifique ouvert d'autant plus urgent que les systèmes gagnent en capacités et en autonomie.

Une autre difficulté majeure réside dans le fait qu'il n'existe pas actuellement de moyen de garantir ni de vérifier formellement l'alignement d'un système d'IA avant son déploiement. On ne peut qu'estimer cette propriété en évaluant les comportements des systèmes avant de les déployer, une approche limitée en premier lieu par l'absence d'évaluation fiable de l'alignement, et plus largement par le fait qu'un système suffisamment capable peut se comporter différemment dans et hors des contextes d'évaluation (par exemple donnant des réponses honnêtes seulement tant qu'il détecte qu'il est évalué)[32]. Les techniques actuelles visant à identifier les systèmes mésalignés sont utiles mais structurellement incomplètes face à ce problème.

Ainsi :

- **L'alignement est un problème scientifique non résolu**. Aucun fournisseur de modèle, y compris ceux qui développent les systèmes les plus avancés, ne dispose aujourd'hui de

[26] "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.
[27] Edwards et Emberson. "Open models lag state-of-the-art closed models by 4 months." Epoch AI, Data Insights (2026). https://epoch.ai/data-insights/open-closed-eci-gap
[28] SecureBio. "Benchmark Trends." SecureBio (2026.). https://securebio.org/benchmarks.
[29] Bengio et al. "International AI Safety Report 2026." (2026). https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[30] Maier, et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." arXiv (2025). https://doi.org/10.48550/arXiv.2510.02840
[31] Lynch et al. "Agentic Misalignment: How LLMs Could Be Insider Threats." arXiv (2025). https://doi.org/10.48550/arXiv.2510.05179
[32] Needham et al. (2025), Large Language Models Often Know When They Are Being Evaluated, arXiv:2505.23836.

méthodes garantissant que des systèmes fortement généraux et très performants n'adoptent aucun comportement mésaligné[33].

- **Les conséquences du mésalignement des IA frontière croissent avec l'autonomie et la portée des actions**[34]. Un système mésaligné confiné à la production de texte dans une conversation pose des risques limités, tandis qu'un système désaligné autonome interagissant avec des acteurs externes peut causer d'importants dégâts.

En ce sens, la catégorie **AGI au seuil critique** (AGI-SC) délimitée dans ce rapport regroupe les systèmes dont les capacités sont telles que leur déploiement en l'absence de mécanismes de contrôle stricts (par exemple, équivalents aux niveaux de certitude requis dans l'industrie aéronautique, agroalimentaire ou nucléaire) poserait des enjeux de sécurité et de contrôle globaux et potentiellement irréversibles.

Résumé de la Partie I

- **Les capacités des IA progressent rapidement** et pourraient atteindre des niveaux équivalents voire supérieurs aux experts humains dans tous les domaines cognitifs.
- **La montée du niveau d'agentivité des systèmes d'IA fait émerger de nouvelles classes de risques**, un système plus agentique étant en mesure d'interagir avec des environnements complexes, de planifier sur le temps long et d'atteindre des objectifs sans supervision humaine. Il n'existe actuellement pas de mesure consensuelle de la généralité d'un système d'IA et de son niveau d'agentivité.
- Une échelle d'agentivité permet de pallier cette lacune, proposant trois classes : les **systèmes faiblement agentiques**, les **systèmes agentiques limités** et les **systèmes AGI au seuil critique (AGI-SC)**.
    - Les GPAI actuelles les plus avancées relèvent du haut de la catégorie « systèmes agentiques limités », et les progrès rapides en termes de capacités pourraient amener à un franchissement du seuil AGI-SC à un horizon qu'il est essentiel d'anticiper si un pays comme la France souhaite prévenir les atteintes à sa sécurité nationale.
    - Les risques suivants sont susceptibles de se manifester selon la catégorie de système :

[33] Bengio et al. "Managing extreme AI risks amid rapid progress." Science 384(6698):842–845 (2024). https://doi.org/10.1126/science.adn0117 ; Greenblatt et al. "Alignment faking in large language models." arXiv (2024). https://doi.org/10.48550/arXiv.2412.14093

[34] Chan et al. "Harms from Increasingly Agentic Algorithmic Systems." FAccT '23, pp. 651–666 (2023). https://doi.org/10.1145/3593013.3594033

Tableau 3 - Synthèse des risques pouvant émerger selon la classe d'agentivité d'un système d'IA.

	ENJEUX SOCIAUX ET ÉCONOMIQUES[35]	MÉSUSAGES[36]	MÉSALIGNEMENT[37]
Systèmes faiblement agentiques	Relations parasociales chez les personnes vulnérables	Désinformation, menace cyber sur cibles d'opportunité	Pertes économiques limitées et dommages réputationnels
Systèmes agentiques limités	Relations parasociales massifiées, disparition de certains secteurs d'emploi, perte d'expertise, concentration de la valeur	Cyberattaques d'ampleur, risque d'augmentation des capacités bioterroristes d'acteurs malveillants	Mésalignement grave pouvant conduire à des pertes économiques et humaines importantes
Systèmes AGI au seuil critique	Concentration extrême du pouvoir technologique et économique, chômage de masse, perte de spécificité du caractère humain	Montée des tensions entre États, automatisation de la R&D en armement, risques de mésusage catastrophiques par des acteurs malveillants	Mésalignement catastrophique, perte de contrôle irréversible à très grande échelle envisageable

Conclusion : Les progrès récents en IA ont permis de concevoir des systèmes plus performants, généraux et agentiques. Le degré d'agentivité d'un système d'IA est défini par sa capacité à atteindre un objectif de haut niveau, en environnement complexe et sans supervision humaine. L'agentivité est une propriété recherchée par les entreprises qui conçoivent les GPAI, en raison des gains économiques générés par des systèmes plus autonomes, mais accroît significativement les risques posés par le déploiement de tels systèmes. Cette note propose de définir un seuil d'agentivité, nommé AGI-SC, qui correspond à un niveau de capacités suffisant pour poser des problèmes de sécurité et de contrôle critiques.

[35] Trammell et Korinek. "Economic Growth under Transformative AI." NBER Working Paper No. 31815 (2023). https://doi.org/10.3386/w31815
[36] Bengio et al. "International AI Safety Report 2026" (2026). https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[37] Kulveit et al. "Position: Humanity Faces Existential Risk from Gradual Disempowerment." Proceedings of the 42nd International Conference on Machine Learning, PMLR 267 (2025). https://proceedings.mlr.press/v267/kulveit25a.html

II. À quel point l'automatisation en cours de la R&D en IA est-elle susceptible d'accélérer l'atteinte des seuils de capacités critiques ?

1. L'automatisation de la R&D implique une accélération rapide des performances et de l'autonomie des systèmes d'IA

A. En quoi consiste l'automatisation de la R&D et quelles en sont les conséquences immédiates ?

L'automatisation de la recherche et développement (R&D) en intelligence artificielle désigne l'utilisation de systèmes d'IA pour accomplir une partie ou l'entièreté des tâches réalisées par les chercheurs et ingénieurs en IA. Ces tâches recouvrent un spectre large d'activités, allant de la génération d'hypothèses à l'écriture et l'exécution de code informatique, en passant par l'analyse de résultats d'expériences[38]. Dans le cas d'une automatisation de la R&D totale, les systèmes d'IA sont en capacité de conduire des programmes de recherche complets sur des horizons longs sans aucune supervision humaine.

Dans tous les domaines économiques, l'automatisation ne tend qu'à augmenter la productivité, mais l'automatisation de la R&D en IA est d'une nature différente. La conception par des systèmes d'IA de systèmes plus avancés, à une vitesse dépassant les capacités de supervision humaine, introduit une dynamique de rétroaction rapide vouée à se propager à tous les autres domaines technologiques[39]. L'automatisation de la R&D en IA est donc considérée comme distincte des autres usages de l'IA, en raison de trois types d'enjeux qui découlent de cette dynamique :

- **Accélération incontrôlée du rythme de déploiement et des capacités.** Si des systèmes automatisés deviennent nettement meilleurs que les humains pour mener la recherche en IA, le rythme auquel de nouvelles générations de systèmes sont conçues peut s'accélérer au-delà de la capacité des institutions à comprendre, évaluer et encadrer ces systèmes[40].
- **Intensification de la course à l'AGI et concentration extrême du pouvoir technologique et économique.** L'automatisation de la R&D en IA tend à avantager massivement les acteurs qui y parviennent en premier, l'accélération de la vitesse de progrès promettant un avantage stratégique décisif sur des rivaux économiques et militaires[41]. La dynamique compétitive entre entreprises et entre États crée une pression à accélérer le déploiement de systèmes automatisant la R&D dans tous les domaines. Cette dynamique de course ne correspond pas nécessairement aux préférences des acteurs individuels pris séparément, mais émerge de leurs interactions stratégiques[42]. Ce problème de coordination est structurellement analogue à d'autres courses historiques (armement nucléaire, armes biologiques) mais la vitesse de progression des capacités des systèmes développés et le degré d'autonomie visé

[38] Denain et al. "Toward an O*NET for AI R&D." Epoch AI, Gradient Updates (2026). https://epoch.ai/gradient-updates/toward-an-onet-for-ai-rnd
[39] Davidson et al. "When Does Automating AI Research Produce Explosive Growth? Feedback Loops in Innovation Networks." NBER Working Paper No. 35155 (2026). https://doi.org/10.3386/w35155
[40] Toner et al. "When AI Builds AI: Findings From a Workshop on Automation of AI R&D." Center for Security and Emerging Technology (CSET) (2026), p. 5-7. https://cset.georgetown.edu/publication/when-ai-builds-ai/
[41] Sytsma. "Decisive Economic Advantage: Modeling the Transition from Temporary First-Mover Leads to Economic Dominance in Artificial General Intelligence." RAND Corporation (2026). https://www.rand.org/pubs/research_reports/RRA4444-1.html
[42] 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

rendent difficile la comparaison avec ces événements antérieurs. La capacité des États à préserver leur autonomie dépend directement de leur capacité à anticiper ces phénomènes.

- **Opacité croissante et perte de contrôle sur des systèmes développés.** Dans le paradigme de développement des systèmes d'IA modernes, il existe une divergence irréductible entre les objectifs souhaités par les développeurs, et ceux internalisés par les modèles. Il n'est pas possible de spécifier ou d'identifier les objectifs internalisés[43] et d'obtenir des garanties de contrôle sur leurs comportements[44]. L'automatisation de la R&D crée un risque d'accroissement de l'écart entre l'intention initiale du développeur et le comportement de l'IA effectivement observé une fois déployée. Un système mésaligné par défaut qui à son tour conçoit, entraîne ou évalue la génération suivante pourrait tendre à transmettre et à amplifier ses propres écarts. **Les systèmes d'IA conçus par d'autres systèmes d'IA seront par défaut moins compréhensibles par les humains qui les déploient.** Or, les systèmes de R&D automatisée ne pourront conduire des travaux complexes sur des horizons longs (par exemple, conception d'une nouvelle génération d'IA) que s'ils sont en mesure de poursuivre leurs objectifs avec une grande autonomie. Garantir que ces objectifs soient effectivement alignés avec les intentions humaines, et que les scénarios de perte de contrôle restent sous des seuils de risques prédéfinis, est un problème scientifique non résolu à ce jour.

B. L'automatisation totale de la R&D en IA est un objectif central des entreprises à la frontière du développement de l'IA

Les performances des GPAI dans le domaine de la programmation sont déjà exploitées par les entreprises qui développent des systèmes conçus pour accélérer leurs cycles de R&D. Leur objectif est de parvenir à automatiser entièrement leurs processus de recherche en IA, en développant des modèles capables de coder et de réaliser des expériences en complète autonomie. Les entreprises d'IA de frontière ont fait de cette automatisation un objectif stratégique explicite, assorti d'échéances précises et soutenu par des engagements financiers d'une ampleur sans précédent dans l'histoire industrielle. OpenAI a ainsi annoncé en octobre 2025 viser la construction d'un « système autonome du niveau d'un chercheur junior » d'ici septembre 2026, puis d'un « système de recherche IA automatisé » d'ici 2028, capable « d'accélérer et d'automatiser une part croissante du processus de recherche lui-même »[45]. Dario Amodei, CEO d'Anthropic, a déclaré en janvier 2026 que l'IA écrivait déjà une grande partie du code chez Anthropic, accélérant substantiellement le développement de la prochaine génération de systèmes, et estimé que la boucle de rétroaction par laquelle « la génération actuelle d'IA construit la suivante» pourrait se concrétiser dans un délai d'un à deux ans[46]. Ces déclarations sont à prendre avec un recul critique nécessaire, mais elles sont adossées à des décisions financières qui en renforcent la crédibilité, avec en particulier des investissements dans les centres de calcul qui devraient dépasser 500 milliards de

[43] Maier et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." arXiv preprint arXiv:2510.02840 (2025). https://arxiv.org/abs/2510.02840

[44] GPAI Policy Lab. "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle." Octobre 2025. https://gpaipolicylab.org/note-2.

[45] Altman et Pachocki. "Built to benefit everyone: our plan." OpenAI (2026). https://openai.com/index/built-to-benefit-everyone-our-plan/

[46] Amodei. "The Adolescence of Technology." darioamodei.com (2026). https://darioamodei.com/essay/the-adolescence-of-technology

dollars en 2026[47]. Or, si la puissance de calcul est essentielle pour entraîner les modèles[48], elle l'est aussi pour permettre d'alimenter un grand nombre d'IA agentiques capables de contribuer aux processus de R&D en IA[49].

Encadré 3 - L'automatisation partielle de la R&D en IA a déjà commencé

Certains systèmes contribuent déjà directement à l'amélioration de l'infrastructure sur laquelle repose le développement des modèles eux-mêmes. *AlphaEvolve*, un agent de codage évolutionnaire développé par Google DeepMind a optimisé les programmes de calcul matriciel utilisés pour l'entraînement de Gemini, obtenant une accélération de 23 % sur ces programmes et une réduction de 1 % du temps total d'entraînement du modèle[50]. *AlphaEvolve* illustre ainsi concrètement la boucle récursive par laquelle l'IA améliore l'infrastructure qui sert à entraîner l'IA elle-même.

La publication en avril 2026 par Anthropic de l'analyse des capacités de Claude Mythos fournit un aperçu de l'état actuel de ces capacités d'automatisation de la R&D. Lors d'évaluations externes, Claude Mythos a redécouvert de manière autonome 4 des 5 étapes clés d'une tâche de recherche en *machine learning* inédite, un résultat qui aurait pris plusieurs jours à une semaine à un ingénieur de recherche expérimenté. Les évaluateurs ont observé que le modèle testait des hypothèses, identifiait efficacement ses erreurs et raisonnait de manière compétente sur un problème complexe, constituant « un bond significatif en termes d'utilité pour la recherche ».[51].

Enfin, le document de référence sur les questions de sécurité de l'IA d'Anthropic annonce explicitement : « il est plausible que dès le début de l'année 2027 nos systèmes d'IA seront en mesure d'accélérer très fortement voire d'automatiser entièrement le travail d'équipes entières de chercheurs experts dans des domaines où un progrès rapide pourrait créer des menaces à la sécurité internationale et/ou des ruptures rapides dans l'équilibre global des puissances (par exemple dans le domaine de l'énergie, de la robotique, de la conception d'armes et dans le développement de l'IA elle-même)[52] ».

Ces informations sont difficiles à vérifier en raison de la faible transparence d'Anthropic sur les capacités de ses systèmes les plus avancés, et de l'absence d'évaluations quantitatives précises, mais elles concordent avec l'évolution récente des capacités et les résultats des modèles analysés dans la suite de ce rapport.

La convergence entre les déclarations d'intention, les engagements financiers, les premiers déploiements opérationnels et ces résultats d'évaluation indique que **l'automatisation de la R&D en IA n'est pas une spéculation théorique mais un programme industriel en cours d'exécution,**

[47] Goldman Sachs Global Institute. "Tracking Trillions: The Assumptions Shaping the Scale of the AI Build-Out." Goldman Sachs (2026). https://www.goldmansachs.com/insights/articles/tracking-trillions-the-assumptions-shaping-scale-of-the-ai-build-out
[48] "Puissance de calcul ; chaîne d'approvisionnement et dépendances stratégiques pour le développement de l'IA." GPAI Policy Lab (2026)
[49] You. "Frontier labs don't use most AI compute (yet)." Epoch AI, Gradient Updates (2026). https://epoch.ai/gradient-updates/frontier-labs-dont-use-most-ai-compute
[50] Novikov et al. "AlphaEvolve: A coding agent for scientific and algorithmic discovery." arXiv (2025). https://doi.org/10.48550/arXiv.2506.13131
[51] Anthropic. "Claude Mythos Preview System Card." Anthropic (2026). https://www.anthropic.com/system-cards
[52] Anthropic. "Frontier Safety Roadmap." Anthropic (2026). https://www.anthropic.com/responsible-scaling-policy/roadmap

dont les premières manifestations concrètes sont déjà mesurables[53]. Une telle automatisation permettrait d'accélérer significativement le développement de modèles encore plus performants et autonomes, eux-mêmes mobilisés pour développer une nouvelle génération de modèles. Ce phénomène donnerait aux entreprises à la frontière les moyens de creuser fortement l'écart avec les autres entreprises ne disposant pas encore de systèmes en mesure d'accélérer fortement leur R&D. **Toute tentative d'anticipation des capacités futures des systèmes d'IA qui ne prend pas en compte les effets de l'automatisation de la R&D tendra à sous-estimer la vitesse de progression des GPAI dans un futur proche.**

2. Modéliser les effets de l'automatisation en cours de la R&D en IA permet de mieux anticiper les capacités futures

A. Les modélisations de l'automatisation de la R&D incluent une ou plusieurs boucles de rétroaction

Le cadre mathématique utilisé pour modéliser les effets de l'automatisation de la R&D en IA sur l'évolution des capacités des GPAI est celui de la littérature sur la croissance économique[54]. Dans ces modèles, la croissance de la productivité (pour l'IA, celle des capacités générales) dépend de l'évolution du nombre de chercheurs et de l'accumulation de connaissances, mais avec des rendements décroissants dans la production d'idées (une fois que l'on a fait des percées théoriques, il est de plus en plus difficile de trouver de nouvelles idées)[55]. Intuitivement, cela signifie que pour maintenir un rythme de progrès constant, il faut augmenter l'effort de recherche afin de compenser la raréfaction des idées.

Dans les modélisations étudiées, l'intuition centrale est la suivante : si l'IA peut automatiser la R&D, elle lève la contrainte de rareté du facteur humain dans la production d'idées. On peut alors utiliser un grand nombre d'instances d'un système d'IA pour contribuer à la R&D, en explorant un très grand nombre de possibilités, accélérant ainsi nettement les progrès. **Cela transforme une trajectoire de croissance régulière en croissance très rapide, à condition que la difficulté à découvrir de nouvelles idées n'augmente pas trop vite**. Ainsi, la frontière des capacités dépend non seulement de paramètres exogènes (puissance de calcul, nombre de chercheurs, données), mais aussi de la capacité des systèmes d'IA à contribuer directement au processus de R&D en IA. Plusieurs travaux ont exploré ce type de modélisations (modèles de croissance semi-endogènes), menés tant par des chercheurs en économie que par des chercheurs en IA. Ci-dessous une représentation simplifiée de la structure générale des modèles d'automatisation de la R&D en IA[56] :

[53] Favaro et Clark. "When AI builds itself." Anthropic (2026). https://www.anthropic.com/institute/recursive-self-improvement

[54] Il s'agit plus particulièrement des modèles de croissance semi-endogènes. Ce cadre a été appliqué à l'IA notamment dans un article commun de Philippe Aghion (économiste français, lauréat du prix Nobel d'économie 2025), Benjamin Jones et Charles Jones : Aghion et al. "Artificial Intelligence and Economic Growth," NBER Working Paper 23928 (2017), https://doi.org/10.3386/w23928

[55] Bloom et al. "Are Ideas Getting Harder to Find?" American Economic Review 110(4): 1104–1144 (2020). https://doi.org/10.1257/aer.20180338

[56] Davidson et al. "Three Types of Intelligence Explosion." Forethought (2025). https://www.forethought.org/research/three-types-of-intelligence-explosion

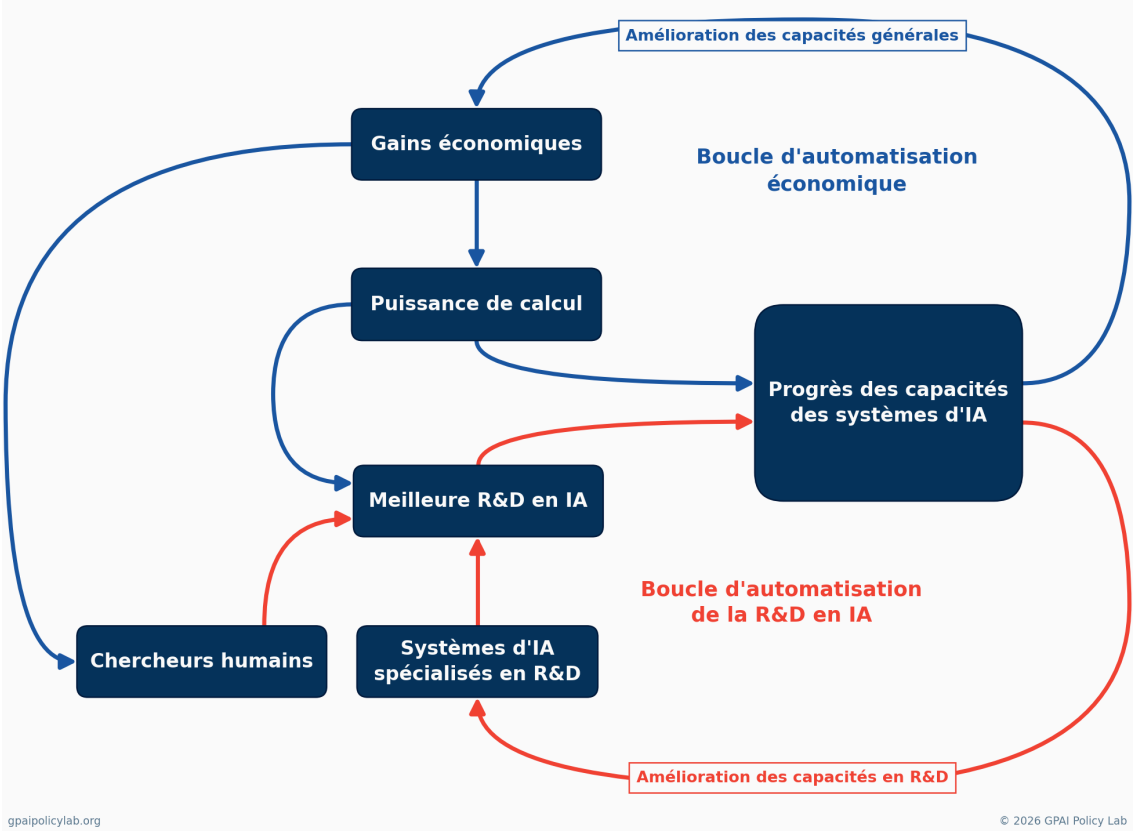

Figure 2 - Schéma simplifié de la structure d'un modèle des effets de l'automatisation de la R&D en IA. On observe deux boucles de rétroaction : en rouge, la boucle qui correspond à une automatisation du processus de R&D via l'augmentation des capacités et du nombre de systèmes d'IA capables d'automatiser partiellement puis intégralement la R&D en IA ; en bleu, la boucle de rétroaction économique qui s'amplifie à mesure que l'automatisation de diverses tâches provoque des gains importants, qui sont ensuite réinvestis pour augmenter encore les capacités des générations suivantes de systèmes d'IA.

La boucle économique est plus lente que la boucle d'automatisation de la R&D, parce qu'elle est contrainte par la capacité à déployer des systèmes capables d'automatiser des tâches de valeur. Il existe une différence importante entre la capacité à automatiser une tâche et l'automatisation effective de cette tâche : le déploiement nécessite des apports en capitaux importants et une adaptation profonde des processus de production pour y intégrer des systèmes d'IA. **À l'inverse, la R&D en IA est une tâche essentiellement cognitive, qui nécessite seulement l'existence de systèmes à très fortes capacités au sein des entreprises à la frontière pour être automatisée. La boucle de rétroaction « automatisation de la R&D » produit donc une contribution plus rapide aux progrès des capacités des GPAI que la boucle « automatisation économique générale »[57].**

[57] Eth et Davidson. "Will AI R&D Automation Cause a Software Intelligence Explosion?" Forethought (2025). https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion

B. L'automatisation de la R&D produit une accélération marquée de la vitesse de progrès

La modélisation des effets sur les capacités générales de l'automatisation de la R&D en IA peut se représenter en trois phases distinctes.

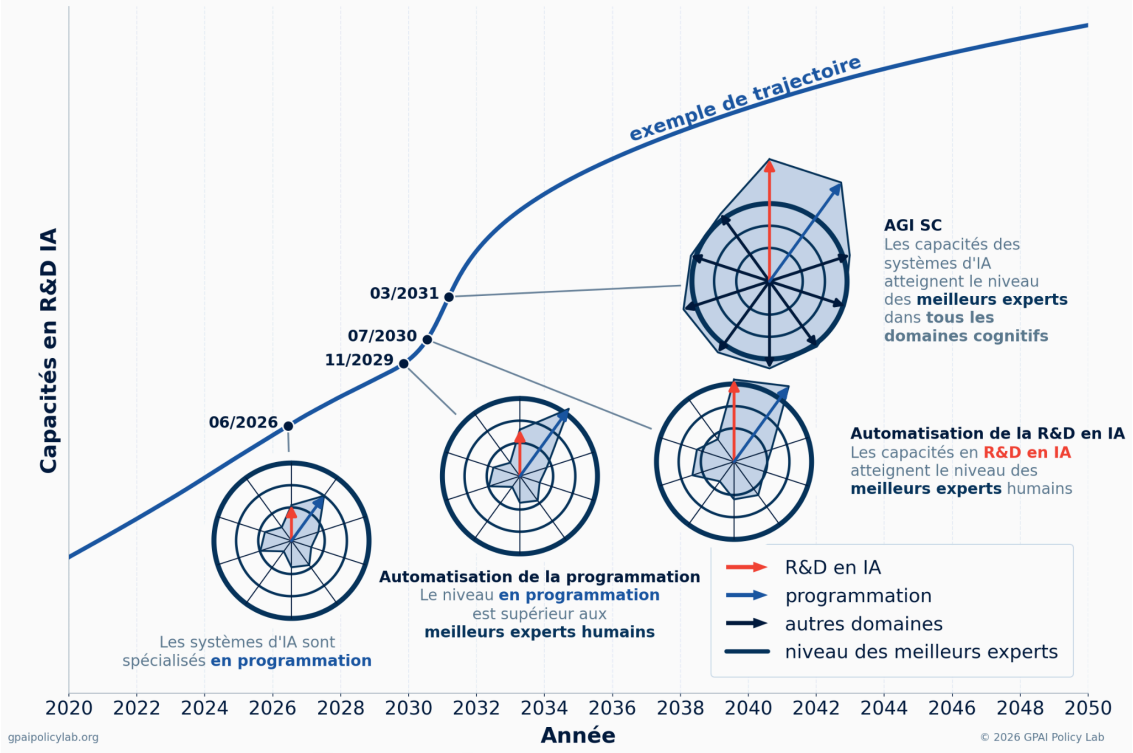

Figure 3 - Schéma simplifié de l'évolution du profil de compétences des systèmes d'IA dans un scénario d'automatisation totale de la R&D en IA. La trajectoire provient du modèle AI Futures (Annexe B), et représente un scénario où le seuil AGI-SC est franchi mi-2031.

- **La première phase correspond à l'amélioration des capacités en programmation informatique** (flèche bleue), domaine qui présente l'avantage d'être facilement vérifiable (le code fonctionne ou non). Cette phase s'arrête lorsque les performances des meilleurs systèmes d'IA égalent et surpassent les experts dans ce domaine précis
- **La deuxième phase est dédiée à l'optimisation des capacités en R&D** (flèche rouge), domaine plus difficile car nécessitant des connaissances transversales et un comportement exploratoire (analyser des idées et identifier leur pertinence avant de les exécuter)
- **La troisième phase est celle de la généralisation du profil de compétences** : lorsque le niveau des systèmes d'IA en R&D a dépassé le niveau des meilleurs chercheurs, l'accélération du progrès est telle que la progression sur l'ensemble des domaines de compétence devient rapide[58]. **C'est cette phase qui correspond au franchissement du**

[58] L'hypothèse est qu'un système aux capacités surhumaines en R&D en IA est en mesure d'optimiser les architectures et l'efficacité algorithmique à un niveau permettant l'acquisition de compétences générales expertes dans tous les domaines, il s'agit donc d'une forme de transition de phase dans les capacités.

seuil AGI-SC, c'est-à-dire à la conception de systèmes très agentiques et dont les performances dépassent les meilleurs experts dans tous les domaines cognitifs. Le profil de compétences devient "homogène" par rapport aux capacités humaines, c'est-à-dire au moins égal aux meilleures performances des experts dans chaque domaine.

Trois cadres de modélisation différents sont retenus ici, afin d'analyser les dynamiques communes entre ces modèles et de vérifier la cohérence de leurs projections. Les modèles suivants ont été choisis pour la complémentarité de leurs approches :

- **le modèle SIE (pour *Software Intelligence Explosion*)**[59] est un modèle qui se concentre uniquement sur la boucle d'automatisation de la R&D en IA et spécifiquement sur la phase post-automatisation totale. [Voir Annexe A pour une description détaillée].
- **le modèle AIF (pour *AI Futures*)**[60] vise à modéliser en détail les déterminants qui conduisent à l'automatisation de la R&D puis la dynamique d'accélération des capacités, tout en tenant compte de la complexité du processus de R&D et des freins qui peuvent limiter l'automatisation de la R&D. [Voir Annexe B pour une description détaillée].
- **le modèle GATE**[61] **(*Growth and AI Transition Endogenous model*)** a la particularité d'être un modèle macro-économique couplé à un modèle de capacités, qui s'intéresse à l'évolution des capacités de l'IA via la boucle de rétroaction économique, c'est à dire via l'automatisation de toutes les tâches économiquement rentables à travers un flux d'investissement. [Voir Annexe C pour une description détaillée].

3. Les trois modèles convergent vers des dynamiques similaires

Pour obtenir une comparaison cohérente, il est nécessaire d'harmoniser les modèles entre eux. Deux méthodes d'harmonisation ont été retenues pour anticiper le franchissement de seuils de capacités et les risques associés : l'une basée sur le facteur d'accélération par rapport au rythme historique, l'autre basée sur la mise en équivalence de seuils de capacités dans les modèles, par rapport à l'échelle d'agentivité proposée dans le Tableau 2 qui définit le seuil AGI-SC.

A. Harmonisation par facteur d'accélération

On peut comparer l'accélération qui se produit lors de l'automatisation de la R&D en IA à la vitesse historique des progrès des capacités. Cette approche présente l'intérêt de quantifier la rapidité à laquelle les capacités des modèles pourront s'améliorer par rapport à ce que l'on a pu observer précédemment, et reprend exactement le cadre développé dans le modèle ***Software Intelligence Explosion (SIE)***. Elle concorde aussi avec l'une des métriques mentionnées dans les doctrines de sécurité des entreprises à la frontière (*responsible scaling policy,* RSP) comme un seuil

[59] Davidson et Houlden. "How quick and big would a software intelligence explosion be?" Forethought (2025). https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be
[60] AI Futures Project. "AI Futures Model." AI Futures Project (2025). https://www.aifuturesmodel.com/
[61] Erdil et al. "GATE: An Integrated Assessment Model for AI Automation." arXiv (2025). https://doi.org/10.48550/arXiv.2503.04941

d'alerte : pour Anthropic, le seuil est défini **comme doublement (x2) de la vitesse de progrès[62], et pour Open AI un quintuplement (x5)[63]**.

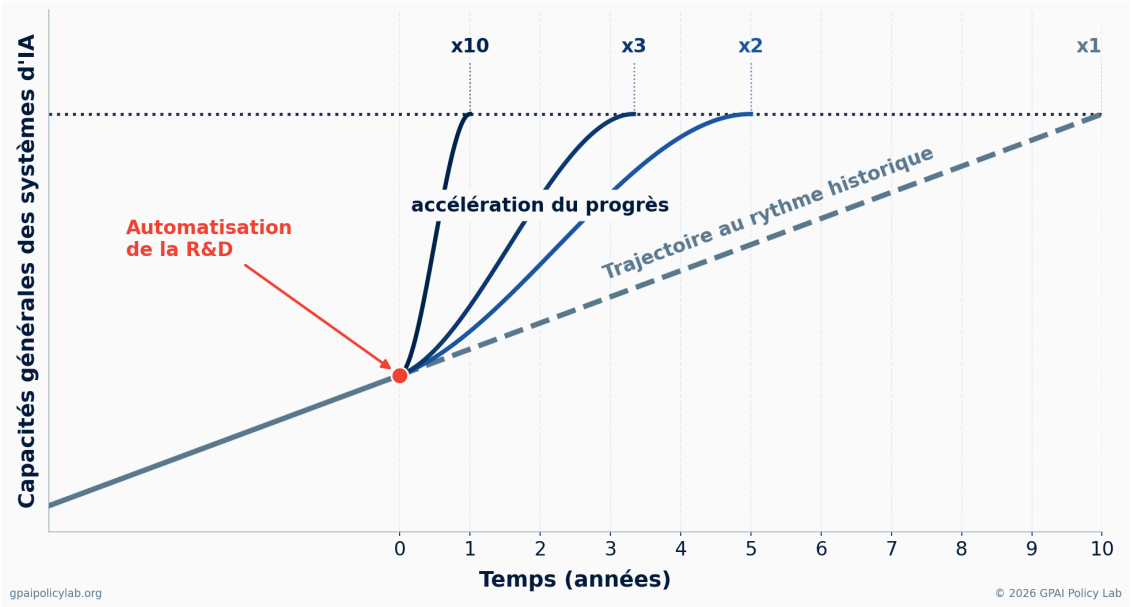

Figure 4 - Trajectoires d'accélération potentielles comparées au rythme historique de progrès en IA. Le seuil en pointillés représente schématiquement un niveau de capacité critique : dans l'hypothèse où le rythme actuel de progrès conduirait à l'atteindre en 10 ans, l'automatisation de la R&D pourrait, selon le facteur d'accélération atteint, réduire ce délai à 5 ans (doublement) ou moins.

On peut utiliser les données du modèle AIF et SIE pour **calculer un facteur d'accélération par rapport aux progrès historiques en IA observés**[64], une fois l'automatisation de la R&D amorcée et obtenir la figure suivante.

[62] Anthropic. "Anthropic's Responsible Scaling Policy (Version 3.3)." Anthropic (2026), p. 10. https://cdn.sanity.io/files/4zrzovbb/website/dd0ec579bee2cd144069c478ede3e35ea080ad02.pdf

[63] OpenAI. "Preparedness Framework (Version 2)." OpenAI (2025), p. 6. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[64] Pour ce faire, on utilise la métrique effective compute (le produit de training compute et de software efficiency), étant donné que le modèle suppose une relation log-linéaire entre les capacités générales d'un système d'IA et la valeur en ordre de grandeur d'effective compute. On utilise la plage 2024.0 - 2025.0 pour AIF, et le rythme historique initial pour SIE.

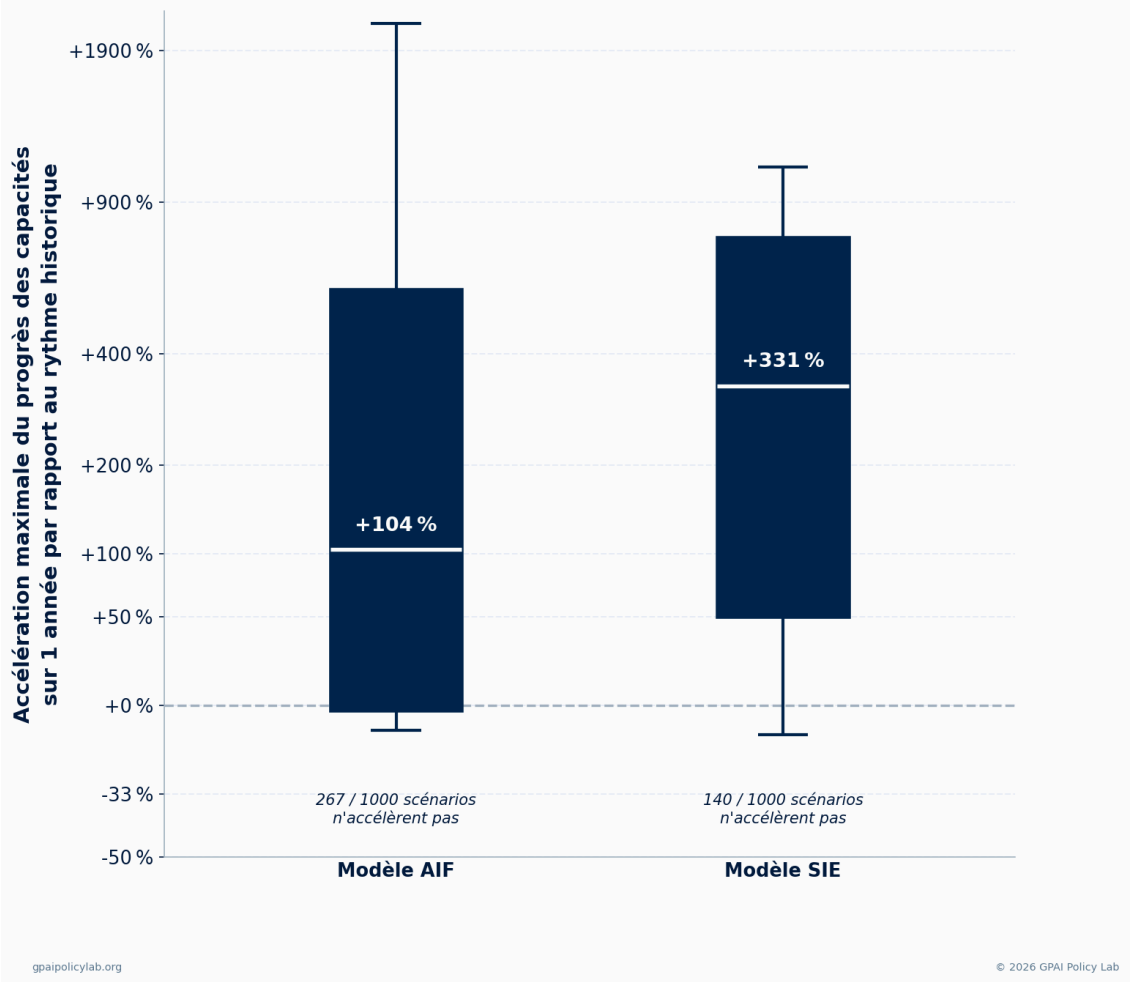

Figure 5 - Comparaison de l'accélération maximale sur 1 an dans le modèle SIE et le modèle AIF. Les boîtes représentent l'intervalle de confiance à 50% et les moustaches à 80%. Le modèle AIF prédit une accélération médiane du niveau de capacités près de deux fois plus rapide (+104%) que le rythme de progrès historique. Les prédictions du modèle SIE sont encore plus agressives (quatre fois plus rapides). Pour comparaison, un quadruplement de la vitesse d'évolution des capacités correspondrait à effectuer un saut équivalent à celui observé entre ChatGPT (2022) et Claude Mythos (2026) réalisé en moins d'un an.

La comparaison des deux modèles en termes de facteur d'accélération permet d'affirmer avec un haut degré de certitude que **la vitesse actuelle des progrès des capacités des systèmes d'IA est inférieure à celle qui sera atteinte dans le futur, en raison de progrès plus rapides attribuables à l'automatisation de la R&D en IA.**

Pour le modèle GATE, la comparaison est difficilement réalisable, parce qu'il s'agit d'un modèle global qui ne distingue pas la R&D en IA des autres tâches à valeur économique[65]. **GATE joue un rôle complémentaire par rapport aux modèles SIE et AIF, en prenant en compte les contraintes**

[65] Pas de boucle de rétroaction sur l'efficacité logicielle autre qu'une boucle économique via l'investissement. Par conséquent, la majorité des progrès sont portés par une croissance rapide de la puissance de calcul et très peu par les progrès logiciels, ce qui rend difficile la comparaison avec les métriques utilisées par SIE et AIF.

macro-économiques à l'investissement qui sont nécessaires à l'augmentation des capacités. La modélisation GATE met en évidence que les trajectoires de capacités décrites par SIE et AIF sont atteignables économiquement. Autrement dit, la dynamique d'accélération des investissements en IA ne bute pas sur une contrainte macroéconomique : les flux de capitaux et le déploiement d'infrastructures de calcul nécessaires à l'automatisation totale du travail restent cohérents avec la logique d'allocation des investissements. GATE apporte des arguments solides contre la thèse du ralentissement des capacités en raison des freins économiques : **même un taux d'automatisation limité (environ 15%) suffit à provoquer une augmentation rapide de la croissance, qui permet un réinvestissement des gains dans des infrastructures permettant une automatisation plus poussée.** Le modèle représente ainsi une croissance rapide de l'investissement dans les infrastructures, qui peut déjà être observée[66]. [Voir Annexe C]

Encadré 4 - Croissance des revenus des entreprises à la frontière

Les revenus annualisés (projection du dernier mois sur douze mois) des entreprises à la frontière ont subi une croissance extrêmement rapide en 3 ans : OpenAI a dépassé 25 milliards de dollars de revenu annualisé début 2026, contre environ 6 milliards en 2024, pour un chiffre d'affaires réel 2025 de l'ordre de 13 milliards ; Anthropic est passé d'environ 9 milliards fin 2025 à 47 milliards annoncés fin mai 2026, puis à 65 milliards en août. Les dépenses totales d'investissement consacrées à l'IA approchent 1 000 milliards de dollars en 2026, soit près d'1 % du PIB mondial. Cette croissance est nourrie par l'investissement mais aussi par l'adoption massive des systèmes d'IA de frontière par les particuliers, les entreprises et les gouvernements. Une telle croissance pourrait être symptomatique d'une bulle spéculative, mais la croissance observée des revenus des entreprises d'IA et les modélisations de GATE tendent à écarter cette hypothèse.

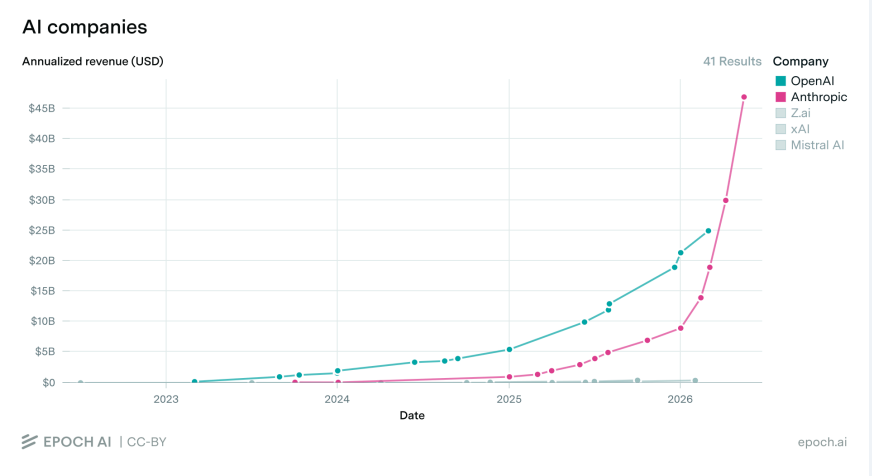

Figure 6 - Évolution du revenu annualisé d'Open AI et Anthropic[67], source : Epoch AI

[66] Juniewicz. "Hyperscaler capex is on trend to outpace their cash inflows by the end of 2026." Epoch AI, Data Insights (2026). https://epoch.ai/data-insights/hyperscaler-capex-vs-cash-flow

[67] Source : https://epoch.ai/data/ai-companies?view=graph&tab=revenue.

Les trois modèles sont cohérents avec un scénario d'accélération rapide des capacités malgré leurs dynamiques distinctes. **Cela signifie qu'il est impossible d'estimer l'évolution des capacités de l'IA en se basant uniquement sur le rythme de progrès historique, car le rythme futur est amené à être nettement supérieur (doublé, triplé, voire plus).** En conséquence, des sauts de capacités rapides pourraient être observés, alors même que les capacités actuelles sont déjà au niveau des experts humains dans certains domaines cognitifs.

B. Harmonisation par seuils de capacités équivalents

La classification par niveau d'agentivité présentée dans le Tableau 2 a permis d'identifier un seuil, nommé AGI-SC, correspondant à des capacités suffisamment élevées pour poser d'importants risques socio-économiques, sécuritaires et de perte de contrôle. Ce seuil peut ensuite être identifié dans chacun des trois modèles, afin d'estimer la date à laquelle il serait franchi étant donné les hypothèses de modélisation.

- Pour le modèle AIF, l'une des bornes du modèle correspond précisément à la définition d'AGI au seuil critique (AGI-SC), ce qui permet d'établir une équivalence directe.
- Pour le modèle GATE, le seuil qui semble le mieux correspondre est le développement d'IA **capables de réaliser l'ensemble des tâches à valeur économique**.
- Pour le modèle SIE, il est plus complexe d'identifier un seuil précis, parce que le modèle ne génère que des trajectoires d'accélération à partir du point d'automatisation totale de la R&D en IA, non défini dans le temps. On peut néanmoins ajouter des hypothèses qui permettent de déterminer une distribution de dates. [Voir Annexe D pour les détails de l'harmonisation]

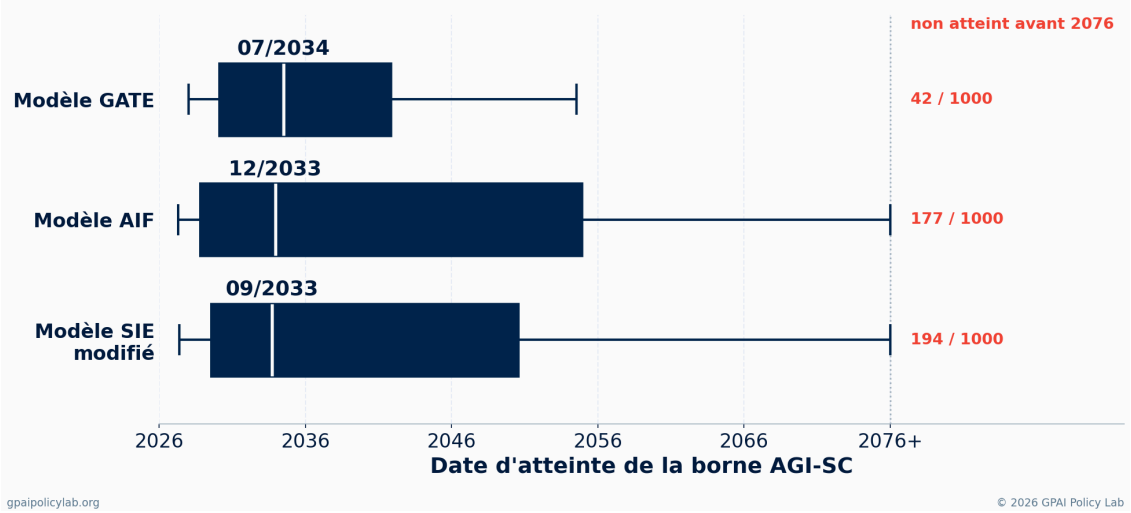

Figure 7 - Harmonisation des trois modèles par date de franchissement de la borne AGI-SC. Tous les franchissements supérieurs à 50 ans (2076 et plus) sont traités comme une seule valeur. (boîtes : IC50, moustaches : IC80, 1000 trajectoires par modèle)

Les résultats harmonisés convergent vers un intervalle temporel compris entre 2030 et 2040, avec une médiane située au plus tard fin 2034 pour le franchissement du seuil AGI-SC, correspondant à un système d'IA fortement agentique, capable de performances supérieures aux meilleurs experts dans tous les domaines cognitifs. Ces dates sont cohérentes avec les choix économiques et les orientations de recherche des entreprises à la frontière, dont l'objectif affirmé est d'atteindre ce niveau de capacités pour automatiser une partie significative des tâches actuellement réalisées par des travailleurs humains.

Ce niveau de capacité présente des risques élevés, tant relatifs aux mésusages et à la déstabilisation économique profonde qu'aux possibilités élevées de perte de contrôle que représentent ces systèmes. Une médiane à 2035 signifie que 50% des scénarios franchissent ce seuil d'ici moins de 10 ans, mais les scénarios les plus rapides (environ 25%) situent ce franchissement avant 2030. L'ampleur et la rapidité des transformations provoquées par des systèmes d'IA disposant d'un tel niveau de capacités sont sans commune mesure avec les précédentes révolutions technologiques[68].

Résumé de la Partie II

- On observe déjà des systèmes d'IA capables d'accélérer les cycles de R&D dans les entreprises à la frontière, et l'automatisation totale est un objectif central dans leur stratégie de développement, soutenue par des investissements massifs.
- L'avancée des performances des GPAI dans tous les domaines est susceptible d'accélérer significativement par rapport au rythme observé actuellement, pouvant doubler voire tripler.
- Les trois cadres de modélisation d'automatisation de la R&D analysés suggèrent que le seuil AGI-SC, présentant des risques élevés, pourrait être franchi avant 2035 (médiane) voire avant 2030 (environ 25%).
- Les données économiques et les choix réalisés par les entreprises à la frontière étayent les trajectoires des différents modèles analysés.

Conclusion : L'automatisation totale de la R&D en IA est activement poursuivie par les entreprises à la frontière, tant par leurs investissements en infrastructures de calcul que par leurs programmes de recherche. Elle est susceptible de provoquer une accélération rapide des capacités de l'IA par rapport au rythme historique observé, ce qui conduirait au franchissement du seuil AGI-SC à horizon 2035, voire avant 2030 dans les trajectoires les plus rapides.

[68] Trammell et Korinek. "Economic Growth under Transformative AI." NBER Working Paper No. 31815 (2023). https://doi.org/10.3386/w31815

III. Quels sont les leviers qui permettraient de ne pas dépasser les seuils critiques avant d'avoir développé des garanties de contrôle ?

Environ 25% des trajectoires simulées par les trois modélisations franchissent le seuil AGI-SC avant 2030, soit dans moins de 4 ans. L'échéance courte avant le franchissement du seuil AGI-SC qui apparaît à travers l'analyse des modèles ne laisse qu'un temps réduit aux interventions visant à limiter les risques liés à l'arrivée de systèmes d'IA très agentiques et performants. Une probabilité d'environ 25% (c'est-à-dire 1 chance sur 4) mérite une attention particulière : **les politiques de défense, de sécurité sanitaire ou de prévention des risques majeurs intègrent systématiquement des scénarios dont la probabilité annuelle est bien inférieure (accidents nucléaires, crues exceptionnelles, pandémies) dès lors que les conséquences potentielles sont d'une gravité suffisante**. Concernant les scénarios de franchissement rapide des seuils de risque, la combinaison d'une probabilité non négligeable et de conséquences majeures pour la souveraineté, la stabilité économique et la sécurité internationale justifie que de telles trajectoires soient traitées comme des scénarios plausibles.

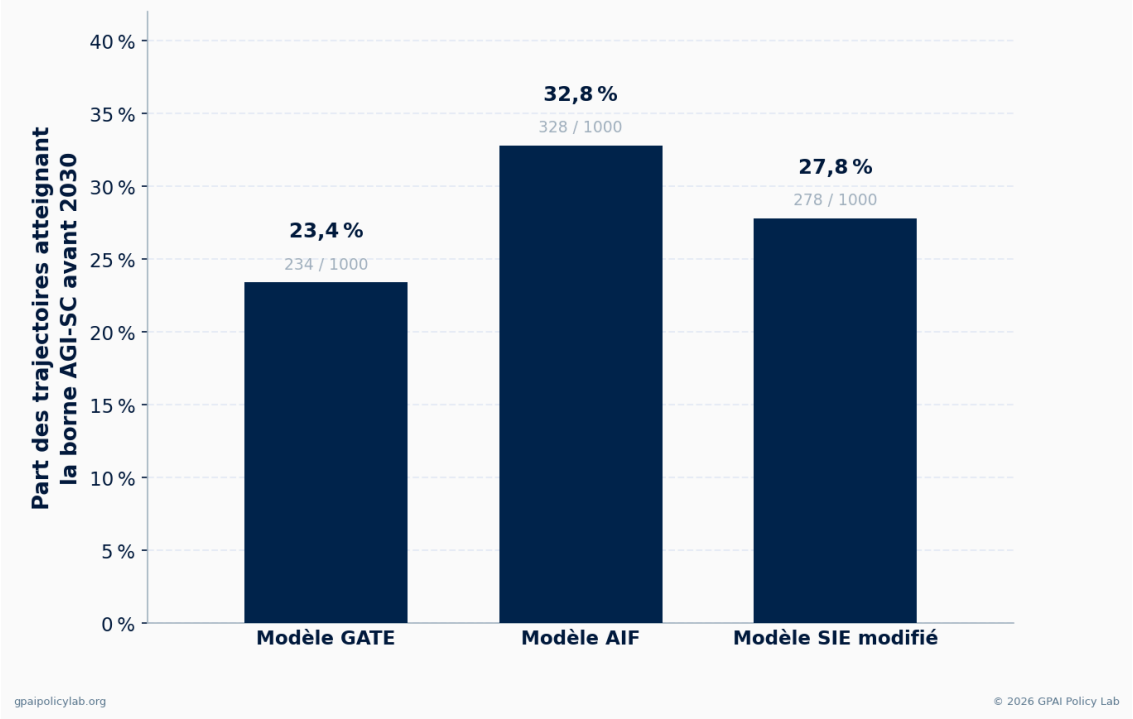

Figure 8 - Proportion de trajectoires simulées par les trois modélisations atteignant AGI-SC avant 2030.

Ces scénarios rapides se distinguent par deux caractéristiques majeures : une croissance ininterrompue de la puissance de calcul disponible et la progression extrêmement rapide des capacités des systèmes d'IA sur les tâches spécifiquement liées à la R&D en IA.

1. Plafonner la puissance de calcul avant 2030 limite le progrès des capacités des systèmes d'IA

La puissance de calcul est l'un des déterminants principaux du progrès des capacités des systèmes d'IA. De fait, le processus de R&D en IA nécessite de tester les potentielles innovations par des expériences complexes qui demandent une grande puissance de calcul[69]. La croissance de la puissance de calcul ayant été très soutenue ces dernières années (doublement de la capacité de calcul mondiale tous les 6 à 7 mois)[70], **un ralentissement net limiterait la possibilité d'automatisation de la R&D en IA et donc l'accroissement très rapide des capacités des systèmes les plus avancés.** Cette dépendance se vérifie dans le modèle AIF, lorsque l'on plafonne la quantité totale de puissance de calcul à un seuil calendaire. **Plus le plafonnement advient tôt, plus le franchissement du seuil AGI-SC est repoussé.**

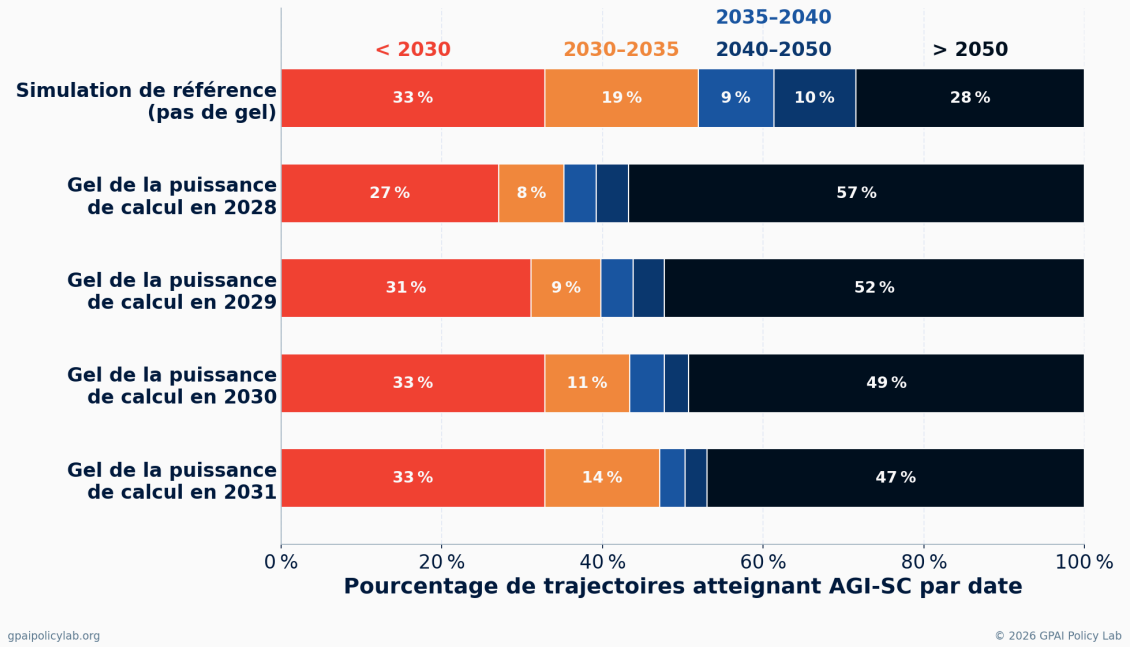

Figure 9 - Évolution de la répartition des dates d'arrivée d'AGI-SC en fonction de la date de gel de la puissance de calcul. Ces trajectoires sont obtenues en forçant la quantité totale de puissance de calcul dans la simulation à rester constante à partir de l'année-seuil. Ainsi, un plafonnement en 2029 réduit de 52 à 40% la proportion de trajectoires franchissant le seuil avant 2035 par rapport au statu quo.

Dans le modèle SIE (*Software Intelligence Explosion*), on considère un scénario simplifié où la puissance de calcul reste constante et où, par conséquent, **les progrès en capacités des systèmes d'IA sont exclusivement attribuables à de meilleurs algorithmes.** En réalité, les entreprises ont déjà investi dans la construction de centres de calcul massifs **qui entreront en fonction entre 2027 et 2030**. On peut donc modifier ce modèle pour mieux tenir compte de cette

[69] Whitfill et Wu. "Will Compute Bottlenecks Prevent an Intelligence Explosion?" arXiv (2025). https://doi.org/10.48550/arXiv.2507.23181
[70] Epoch AI. "Trends in Artificial Intelligence." Epoch AI (2026). https://epoch.ai/trends

complémentarité documentée entre puissance de calcul et progrès logiciel[71], ce qui produit une forte réduction des trajectoires rapides dans le modèle SIE [voir Annexe A].

Le modèle GATE n'ayant pas de boucle de rétroaction sur l'efficacité logicielle (pas de mécanisme explicite d'automatisation de la R&D), **la majorité de la contribution à l'automatisation dans le modèle GATE est due à l'accroissement de la puissance de calcul** (seulement 20 à 30% du chemin pour atteindre AGI-SC est attribuable au progrès logiciel[72], les 70-80% restants étant l'augmentation de la puissance des centres de calcul). Ainsi, les seuils requis pour atteindre l'automatisation totale de l'économie sont inatteignables sans croissance de la puissance de calcul[73].

Ces analyses indiquent que, dans le paradigme actuel d'entraînement des systèmes d'IA, la puissance de calcul est un facteur déterminant. Ce constat est cohérent avec la position actuelle des différents États dans la course à l'intelligence artificielle : les États-Unis, qui développent les modèles les plus avancés possèdent 86% de la puissance de calcul ; la Chine, dont les modèles se situent entre 4 et 10 mois derrière la frontière[74] en possède 8%, et le reste du monde ne dispose pas de modèles à moins d'un an de la frontière[75].

2. Les scénarios sans automatisation rapide de la R&D en IA évitent les trajectoires d'accélération les plus extrêmes

A. La vitesse de progression des capacités des systèmes d'IA sur des tâches de R&D est un paramètre critique

La vitesse de progression des systèmes d'IA sur les tâches de R&D complexes détermine la vitesse d'automatisation totale de la R&D en IA et donc l'accélération du progrès des capacités. En effet, si la performance des meilleures GPAI est déjà équivalente voire supérieure à celle des experts pour des tâches de programmation qui sont facilement vérifiables[76], ce n'est pas le cas pour les tâches nécessitant une bonne intuition (c'est-à-dire la capacité d'évaluer a priori les idées prometteuses ou de définir un programme de recherche). Or, pour parvenir à une automatisation totale de la R&D en IA, il est nécessaire de maîtriser de bout en bout le processus de recherche. Ainsi, les modélisations tentent d'estimer la vitesse à laquelle ces compétences encore peu maîtrisées vont progresser dans les mois et années à venir. Dans la figure ci-dessous, on distingue clairement la différence de régime de progression : la date médiane de franchissement du seuil AGI-SC est particulièrement variable en fonction du groupe (2030 pour le groupe rapide, 2035 pour le groupe modéré, environ 2070 pour le groupe lent).

[71] Whitfill et Wu, op. cit.

[72] Cette proportion est de 100% par construction dans le modèle SIE et d'environ 70-90% dans le modèle AIF

[73] Cela s'explique par la définition du progrès algorithmique qui est choisie par les auteurs du modèle, qui ne prend en compte que les gains d'efficacité (performance fixée, puissance de calcul variable) et non les gains de performance (par ex, à puissance de calcul égal, performances meilleures)

[74] Edwards et Emberson. "Open models lag state-of-the-art closed models by 4 months." Epoch AI, Data Insights (2026). https://epoch.ai/data-insights/open-closed-eci-gap

[75] "Puissance de calcul ; chaîne d'approvisionnement et dépendances stratégiques pour le développement de l'IA." GPAI Policy Lab (2026)

[76] Adamczewski et al. (2026), "MirrorCode: Evidence that AI can already do some weeks-long coding tasks". Published online at epoch.ai. Retrieved from 'https://epoch.ai/publications/mirrorcode-preliminary-results' (2026).

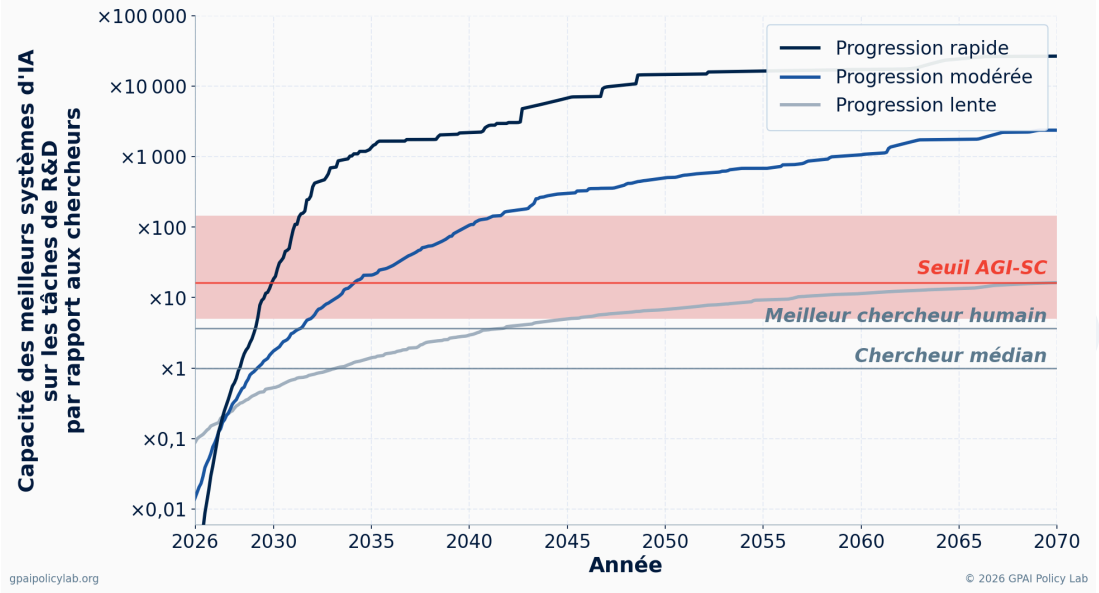

Figure 10 - Décomposition de la progression des capacités des systèmes d'IA en R&D par rapport au niveau des chercheurs humains dans le modèle AIF. Dans cette figure, on distingue trois groupes de scénarios qui ont été isolés par rapport à la valeur du paramètre « progression des capacités en R&D des systèmes d'IA »[77].

On peut s'assurer du caractère critique de ce paramètre en modifiant le modèle AIF pour plafonner la progression des capacités de l'IA en R&D (les seules ressources contribuant au progrès global des capacités deviennent donc l'accroissement de la puissance de calcul et du nombre de chercheurs, ce qui correspond au régime actuel).

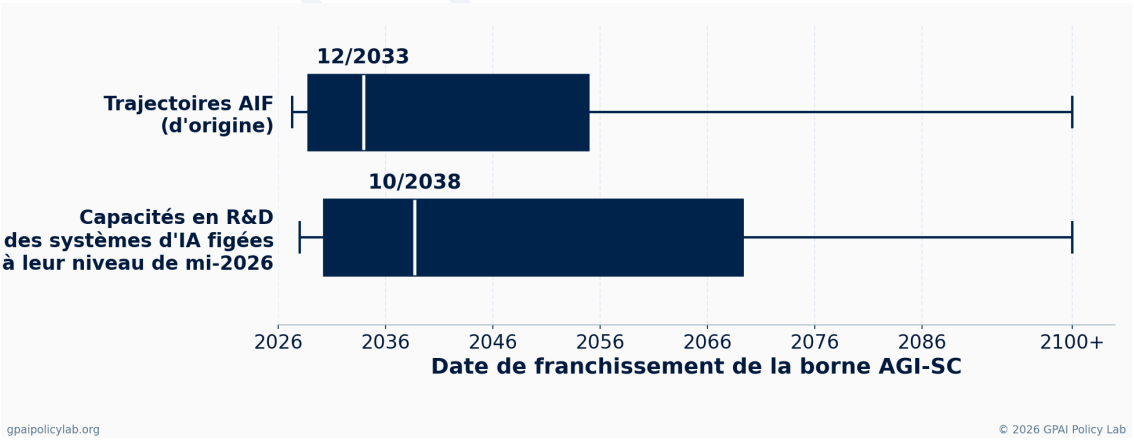

Figure 11 - Comparaison entre les trajectoires classiques (pas d'intervention) du modèle AIF et les trajectoires avec plafonnement des capacités des IA utilisées en R&D à mi-2026. Les boîtes représentent l'intervalle de confiance à 50% et les moustaches à 80%. Le gel correspond à un plafonnement des capacités des IA employées pour accélérer la R&D en IA à leur niveau de mi-2026, en conservant toutes les autres variables non contraintes [Voir Annexe B].

[77] Ce paramètre est défini en SD/OOM de puissance de calcul effective, et la décomposition a été faite par tercile [Voir Annexe technique]

Si les capacités des systèmes d'IA autorisés sur des tâches de R&D sont figées à leur niveau actuel (mi-2026), alors le modèle AIF anticipe un décalage de 5 ans pour l'atteinte du seuil AGI-SC (valeur médiane) et la proportion de trajectoires très rapides (AGI-SC avant 2030) passe de 32,8% à 23%. De façon générale, le gel des capacités en R&D de l'IA au niveau de mi-2026[78] retarde nettement le franchissement du seuil AGI-SC par effet de stagnation de l'effort de recherche en IA, qui reste entièrement limité par les capacités des meilleurs chercheurs humains.

Pour le modèle GATE, il est plus complexe d'analyser les conséquences d'un progrès rapide des capacités des systèmes d'IA dans le domaine de la R&D en IA, parce que le modèle ne possède pas de boucle de rétroaction de ce type (les progrès en R&D sont uniquement mus par les investissements et ne sont pas accélérés à mesure que le pourcentage d'automatisation progresse). On peut cependant modifier la structure du modèle pour simuler les effets d'une telle automatisation, [voir Annexe C] et observer les effets sur la dynamique d'atteinte du seuil AGI-SC.

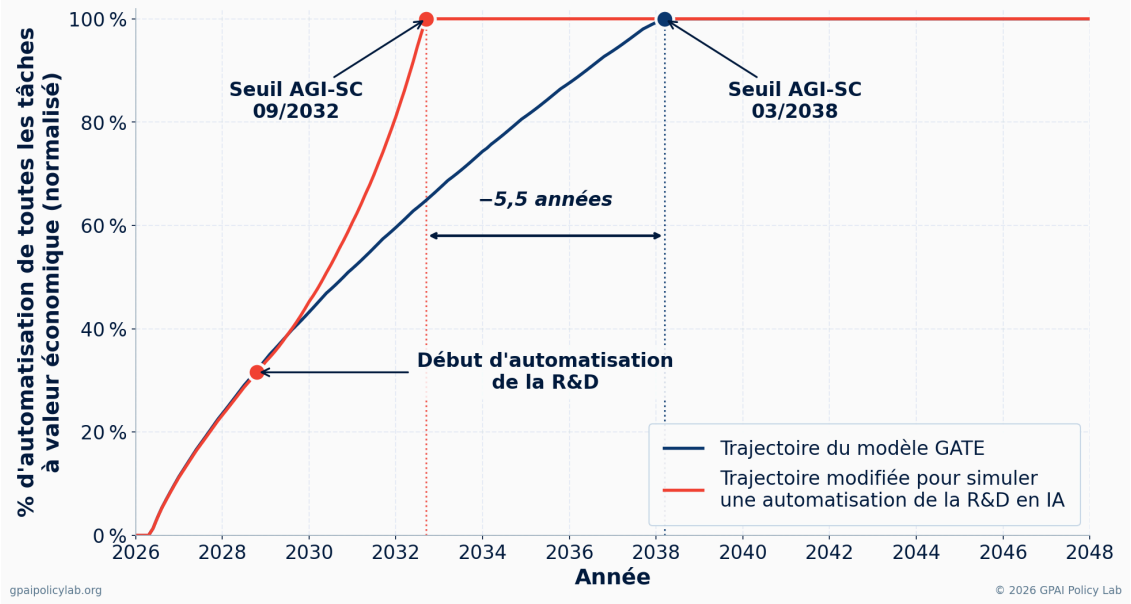

Figure 12 - Comparaison d'une trajectoire standard du modèle GATE avec une trajectoire modifiée pour introduire une boucle de rétroaction directe sur la R&D en IA. La trajectoire d'origine appartient au p75 pour la date AGI-SC.

Ainsi, l'ajout d'une boucle de rétroaction dans le modèle GATE similaire à celle des deux autres modèles étudiés produit une accélération nette du franchissement du seuil AGI-SC. Mais en l'absence de cette boucle, le seuil est également franchi, ce qui signifie que la boucle de rétroaction économique est suffisante pour permettre un progrès des capacités des systèmes d'IA au-delà des seuils critiques.

[78] Gel réalisé via la métrique effective compute

B. Il existe des métriques qui peuvent être suivies pour évaluer la progression des capacités de l'IA en R&D

La vitesse de progression des systèmes d'IA actuels est difficile à estimer directement, car les entreprises à la frontière considèrent l'évolution de leurs capacités internes en matière de R&D automatisée comme une information stratégiquement sensible. Ainsi, les données nécessaires pour calibrer les modélisations présentées dans ce rapport ne sont pas directement accessibles et doivent être estimées à partir d'indicateurs incomplets (performances sur des benchmarks, documents publiés par les entreprises, témoignages divers).

Les *safety frameworks*[79] des entreprises à la frontière (Anthropic, OpenAI et Google DeepMind) témoignent d'une considération de ce risque, mais leurs seuils de déclenchement sont suffisamment ambigus pour limiter le déclenchement de mesures de prévention des risques[80]. De fait, les définitions qui y sont mobilisées partagent une limite commune : elles fixent des seuils d'alerte sans spécifier les protocoles de mesure permettant de déterminer empiriquement si et quand ces seuils sont franchis. Elles ne prévoient pas non plus d'obligation de déclaration à un niveau de granularité suffisant pour permettre une vérification externe, étant donné la criticité de l'avantage comparatif pour une entreprise à la frontière de disposer d'un modèle capable d'automatiser une part importante de la R&D en IA. La sortie du modèle Claude Mythos d'Anthropic illustre les limites rencontrées par cette approche : le système a dépassé les performances humaines sur certaines mesures internes conçues pour évaluer les capacités en R&D, ce qui a conduit Anthropic à conclure au caractère obsolète de ce type de mesures[81]. Pour autant, les compétences générales en R&D et notamment sur les tâches non vérifiables sont encore estimées inférieures à celles des chercheurs humains[82]. Malgré ces limites, les chercheurs d'Anthropic ont estimé que le modèle constitue « un véritable progrès en termes d'utilité pour la recherche en conditions réelles ».

Ainsi, les entreprises à la frontière limitent la publication d'informations relatives aux capacités en R&D de leurs modèles les plus avancés, du fait de difficultés à mesurer effectivement de tels progrès et en raison du caractère stratégique de ces informations. En l'absence d'accès direct à ces données, il existe cependant plusieurs indicateurs qui permettraient d'obtenir une approximation correcte de la vitesse à laquelle les systèmes d'IA progressent en R&D. Parmi eux, les plus opérationnels sont[83] :

- Les performances des systèmes d'IA sur des évaluations spécifiquement conçues pour les tâches de R&D, en incluant des comparaisons avec des experts humains,
- La répartition de l'usage de la puissance de calcul entre utilisation interne et déploiement externe dans la mesure où un glissement vers l'inférence interne signalerait une utilisation croissante de systèmes d'IA au sein du processus de R&D lui-même,

[79] Document public par lequel un développeur de modèles d'IA de pointe s'engage sur la manière dont il gère les risques majeurs liés à ses systèmes.
[80] "Frontier AI Safety Frameworks : Approches et limites face aux scénarios de loss of control" *GPAI Policy Lab* (2025). https://gpaipolicylab.org/note-3.
[81] GPAI Policy Lab. "Claude Mythos Preview : Analyse de l'annonce publique du modèle par Anthropic." GPAI Policy Lab (2026). https://gpaipolicylab.org/blog-2-claude-mythos-preview-analyse.pdf
[82] Anthropic. "Claude Mythos Preview System Card." Anthropic (2026). p33-46. https://www.anthropic.com/system-cards
[83] D'après Chan et al. "Measuring AI R&D Automation." arXiv (2026). https://arxiv.org/abs/2603.03992

- La part du capital investi dans les dépenses de R&D par rapport à la masse salariale, qui constitue un signal de la substitution progressive du travail humain par de la puissance de calcul,
- Le suivi de l'allocation du temps des chercheurs entre tâches primaires de R&D et activités de supervision des sorties produites par des systèmes automatisés pour documenter directement la reconfiguration en cours des flux de travail,
- L'échantillonnage fréquent des avis des chercheurs dans les entreprises à la frontière afin de garder un suivi qualitatif des progrès perçus.

L'agrégation de ces indicateurs permet de limiter l'opacité sur la vitesse à laquelle les systèmes progressent en R&D, mais leur collecte repose pour l'instant essentiellement sur la coopération des entreprises concernées. Par ailleurs, disposer de ces indicateurs permettrait d'estimer le degré d'automatisation en cours de la R&D en IA, mais sans nécessairement présager de la rapidité de l'automatisation future, d'où une nécessité de mettre régulièrement à jour ces métriques afin de pouvoir effectuer un diagnostic précis du type de trajectoire plausible.

Une automatisation rapide de la R&D en IA réduirait la latitude dont disposent les institutions et les États pour peser sur les trajectoires de progression des capacités. Les cycles de décision publique s'étendent sur plusieurs années, tandis que le développement et déploiement de nouvelles générations de systèmes d'IA ne dure que quelques mois aujourd'hui (42 jours entre Claude Opus 4.7 et Opus 4.8), et pourrait se réduire davantage dans un scénario de R&D fortement automatisée où chaque génération de modèle participe à la conception de la génération suivante. Cette accélération signifie qu'au-delà d'un certain rythme, les instruments ordinaires de gouvernance deviennent inopérants. Les décisions prises aujourd'hui sont donc déterminantes dans l'hypothèse d'un progrès rapide. De fait, les conditions dans lesquelles la progression des capacités des systèmes d'IA s'accélère (concentration du calcul, faible transparence sur les capacités, absence de mécanismes internationaux de coordination, course compétitive non régulée entre entreprises) sont dépendantes de choix politiques. **Négliger les scénarios rapides revient donc à sous-estimer le risque et à se priver des leviers permettant d'y faire face.**

Résumé de la Partie III

- **Une automatisation rapide de la R&D en IA provoquerait une augmentation rapide des capacités des meilleurs systèmes d'IA** développés par les entreprises américaines, qui disposent déjà d'une avance considérable en raison de la concentration de la puissance de calcul et de l'automatisation déjà engagée d'une partie des processus de R&D en IA.
- **Dans plus de 25% des scénarios modélisés le seuil AGI-SC est atteint avant 2030**, limitant la fenêtre d'action à moins de 4 années.

- Le franchissement du seuil de capacités AGI-SC accentuerait significativement les enjeux actuellement posés par les systèmes d'IA de frontière, que ce soit en termes de perturbations sociales, économiques, sécuritaires ou de perte de contrôle.
- **Il existe des leviers qui permettent de ralentir le franchissement du seuil AGI-SC**, parmi lesquels :
    - **Le contrôle de la puissance de calcul**, qui est l'un des déterminants majeurs de l'évolution des capacités des systèmes d'IA, à la fois en permettant l'entraînement de modèles plus performants et en étant essentiel à l'automatisation rapide de la R&D.
    - **Le plafonnement des capacités utilisées pour automatiser la R&D en IA.** Le manque de transparence des entreprises à la frontière rend actuellement difficile l'évaluation des capacités des meilleures GPAI actuelles sur les tâches nécessaires pour cette automatisation, mais plusieurs indicateurs permettraient, s'ils étaient collectés, de mieux anticiper la trajectoire future des progrès.

Conclusion

L'automatisation de la R&D en IA est déjà en cours. Les systèmes d'IA actuels accomplissent une part significative des tâches de programmation pour la recherche en IA, et les entreprises développant les IA de frontière orientent explicitement leurs investissements vers l'amplification de cette automatisation.

- L'agentivité des systèmes d'IA les plus performants augmente déjà de génération en génération et recouvre un ensemble de risques socio-économiques, sécuritaires et de perte de contrôle.
- Le seuil AGI-SC (AGI au seuil critique) est défini comme un **niveau d'agentivité permettant à un système d'IA d'acquérir de manière autonome des connaissances et compétences suffisantes pour dépasser les performances des meilleurs experts sur l'ensemble des tâches cognitives.**
- **L'automatisation partielle de la R&D en IA actuellement en cours provoque une accélération mesurable du rythme de progrès des systèmes d'IA à usage général (GPAI) de frontière**. L'année 2026 a vu apparaître des systèmes capables de mener des tâches qui nécessiteraient **plusieurs dizaines d'heures** pour être complétées par des experts, dans des domaines spécifiques.
- L'automatisation totale de la R&D en IA provoquerait une forte accélération des capacités générales des systèmes d'IA, rendant toute anticipation de trajectoire à partir du rythme historique de progrès caduque.
- On peut modéliser les effets de cette automatisation totale sur le rythme de progrès afin d'anticiper les dates plausibles de franchissement du seuil AGI-SC. Les trois modèles

analysés concluent, sous leurs hypothèses initiales et en cas de statu quo dans le domaine de la gouvernance, à une probabilité de franchissement supérieure à 50% avant 2035 et supérieure à 25% avant 2030.

- Il est impossible à l'heure actuelle (mi-2026) d'écarter les scénarios les plus rapides dans lesquels le seuil AGI-SC est franchi avant 2030.
- La modélisation des frictions économiques n'apporte pas d'éléments probants contre la possibilité d'automatiser partiellement voire totalement les tâches actuellement réalisées par les travailleurs humains en une décennie ou moins.

Ces conclusions impliquent des bouleversements difficiles à anticiper à l'heure actuelle. On peut néanmoins supposer que les rapports de puissance entre les États, les dynamiques économiques mondiales et les structures sociales seront profondément altérés à court (2-5 ans) ou moyen (5-10 ans) terme. Les risques associés à l'apparition de systèmes généraux capables de performances supérieures à celles des meilleurs experts sont particulièrement élevés : d'une part en raison d'une capacité à creuser les écarts de puissance entre États et organisations infra-étatiques, d'autre part en raison du caractère incontrôlable dans le paradigme actuel de tels systèmes.

Deux facteurs sont particulièrement déterminants pour la trajectoire des capacités : la puissance de calcul disponible et la vitesse de progression des meilleures IA de frontière dans le domaine de la R&D en IA. On dispose à l'heure actuelle d'une bonne estimation de la répartition de la puissance de calcul, mais l'évolution des performances en R&D reste opaque.

Si la France souhaite peser sur cette trajectoire, la fenêtre pendant laquelle cela reste possible se referme à mesure que les capacités progressent et que les cycles de développement se compressent. Dans l'hypothèse d'une accélération rapide des progrès en IA, le seuil de capacités AGI-SC serait franchi avant 2030, laissant un délai extrêmement réduit pour agir.

Suite possible à cette note

1. Comment mesurer précisément l'évolution de l'agentivité des systèmes d'IA et le risque de perte de contrôle associé ?
2. Quelles sont les données qui permettraient de suivre avec précision le niveau d'automatisation de la R&D en IA atteint par les entreprises à la frontière ?
3. Comment modéliser économiquement les effets d'un plafonnement du niveau d'agentivité des systèmes d'IA, en tenant compte du degré d'automatisation et des arbitrages entre bénéfices et risques ?

Annexes

Le code de toutes les figures de cette note ayant nécessité des données est consultable sur le Github du GPAI Policy Lab au lien suivant : https://github.com/General-Purpose-AI-Policy-Lab

Le code du modèle SIE est public et accessible au lien suivant : https://github.com/thoulden/Accelerated_AI_Progress

Les détails de conception du modèle sont accessibles sur cette page : https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be

Le code du modèle GATE a été transmis par l'un des auteurs (Anson Ho) en avril 2026, avec notification explicite d'usage privé. Ainsi, seules les données ayant servi à générer les figures sont accessibles sur Github. Une sandbox interactive du modèle est disponible ici : https://epoch.ai/gate

Idem pour le code du modèle AIF, transmis par Eli Lifland en juin 2026. Le projet ayant évolué entre-temps, la version la plus récente accessible ici : https://www.aifuturesmodel.com/ diffère du code utilisé pour cette note.

Préambule : Le cadre des modèles de croissance semi-endogènes

Pour étudier les déterminants de la croissance économique sur le temps long, un cadre de modélisation a été posé par Chad Jones en 1995 : celui des modèles de croissance semi-endogènes. À long terme, le taux de croissance du progrès technique y est proportionnel au taux de croissance de l'effort de recherche (c'est à dire la croissance de la population de chercheurs) et non proportionnel au niveau de l'effort de recherche (nombre total de chercheurs), ce qui constitue la différence principale avec les modèles endogènes à la Romer 1990 ou Aghion-Howitt 1992. Formellement, la variation du progrès technique P dans le temps vaut :

$$\frac{dP}{dt} = R^{\lambda} * P^{\Phi}$$

où R est l'effort de recherche (proportionnel à la quantité totale de chercheurs), λ compris entre 0 et 1 est la capacité de parallélisation du travail (doubler le nombre d'individus sur un projet ne double pas nécessairement la vitesse d'avancement), et ϕ détermine la "raréfaction" des idées à mesure que l'on effectue des découvertes. La valeur de ϕ ouvre trois régimes :

- si $\phi > 1$, les rendements du progrès déjà accumulé permettent une boucle de rétroaction rapide : même à effort de recherche constant (R stable), P tend vers l'infini en temps fini (singularité)
- si $\phi = 1$, on retombe sur un modèle purement endogène, où le taux de croissance du progrès est proportionnel au niveau de l'effort de recherche, avec une atténuation via les contraintes de parallélisation

- si $\phi < 1$, à effort de recherche constant, la croissance du progrès ralentit. Pour maintenir la croissance du progrès, il faut que R augmente exponentiellement, ce qui est impossible car la quantité totale de chercheurs est contrainte par la démographie.

Dans le cadre de la R&D en IA, l'hypothèse repose sur le fait que P (le niveau de progrès, exprimé par exemple en termes de capacités des meilleurs systèmes frontière) peut également faire croître R, qui est l'effort de recherche. En effet si les systèmes les plus avancés peuvent se substituer aux chercheurs, soit partiellement soit totalement, alors un P plus élevé implique un R plus élevé aussi, créant une boucle de rétroaction qui peut engendrer une croissance super-exponentielle même sous l'hypothèse $\phi < 1$.

Les trois modèles analysés dans cette note se fondent tous sur ce cadre de modélisation, tout en tenant compte d'autres contraintes, par exemple en traitant R comme une fonction à plusieurs variables.

A. Détails du modèle SIE

Une manière de traduire l'accélération des capacités après automatisation totale de la R&D en IA en termes concrets est de **calculer le nombre d'années de progrès au rythme historique (l'écart observé entre les générations présentes et passées de modèles) que l'on peut réaliser en une seule année**. Ainsi, environ 1 an sépare Claude Opus 4 (mai 2025) de Claude Mythos (avril 2026), et l'écart de capacités entre les deux modèles est très élevé (horizon temporel METR à 50% passant de 1h40 à plus de 16h)[84]. Durant la phase d'accélération, un bond de capacités équivalent pourrait être réalisé en moins de quatre mois si le rythme de progrès triple (facteur d'accélération x3) par rapport au rythme observé historiquement, ce qui conduirait à un franchissement bien plus rapide des seuils de capacités.

Le modèle SIE (Software Intelligence Explosion)[85] tente d'estimer la plage de valeurs de ce facteur d'accélération, en simplifiant au maximum les paramètres qui sont à l'origine de l'amélioration des capacités pour se concentrer uniquement sur le progrès algorithmique, à partir du seuil où l'automatisation de la R&D est entièrement accomplie[86]. **Le modèle postule également une puissance de calcul constante, c'est-à-dire que la rétroaction a lieu essentiellement via la capacité à augmenter le nombre d'agents pour une quantité fixe d'inférence (meilleure efficacité algorithmique), d'où le nom de "Software only intelligence explosion".**

[84] On peut également utiliser la métrique de capacités globales définie par l'Epoch Capabilities Index (ECI)
[85] Davidson et Houlden, op. cit.
[86] Ce seuil, nommé ASARA (pour *AI Systems Automating R&D in AI*), est défini comme celui où une IA pourra réaliser l'équivalent du travail de 30 chercheurs, à une vitesse 30x supérieure.

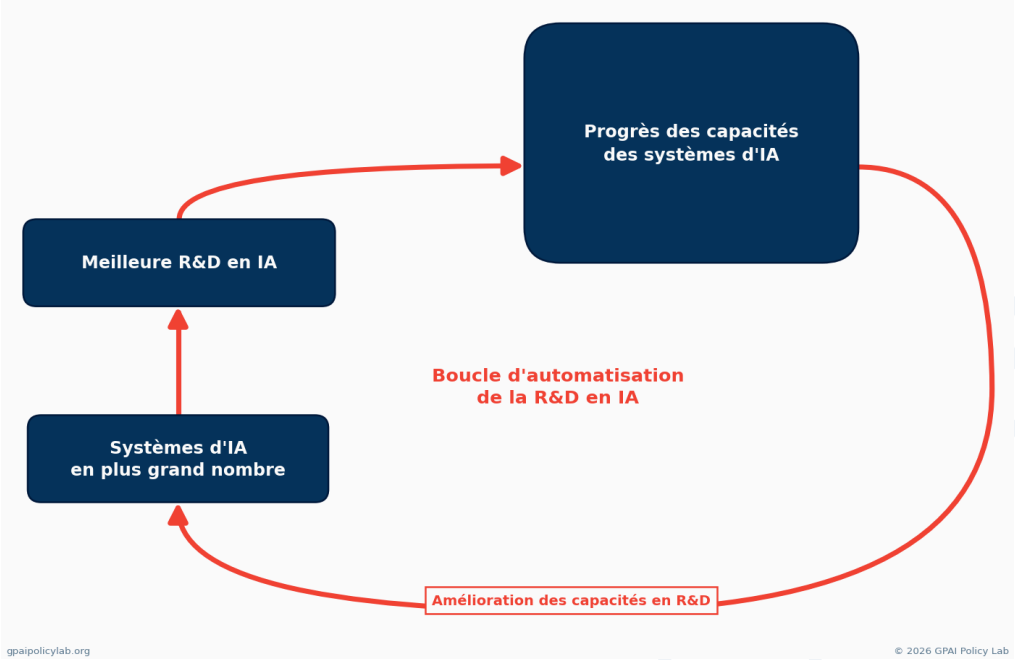

Figure 13 - Schéma simplifié du modèle SIE. En rouge, la boucle de rétroaction du modèle (automatisation de la R&D en IA par l'amélioration des capacités)

Le modèle simule un grand nombre de trajectoires à partir des distributions de paramètres (méthode de Monte-Carlo) pour produire une estimation moyenne en tenant compte des incertitudes fortes qui existent dans le choix des valeurs initiales. Les résultats sur le facteur d'accélération maximal attendu sur une année après l'automatisation totale sont les suivants (paramètres d'origine) :

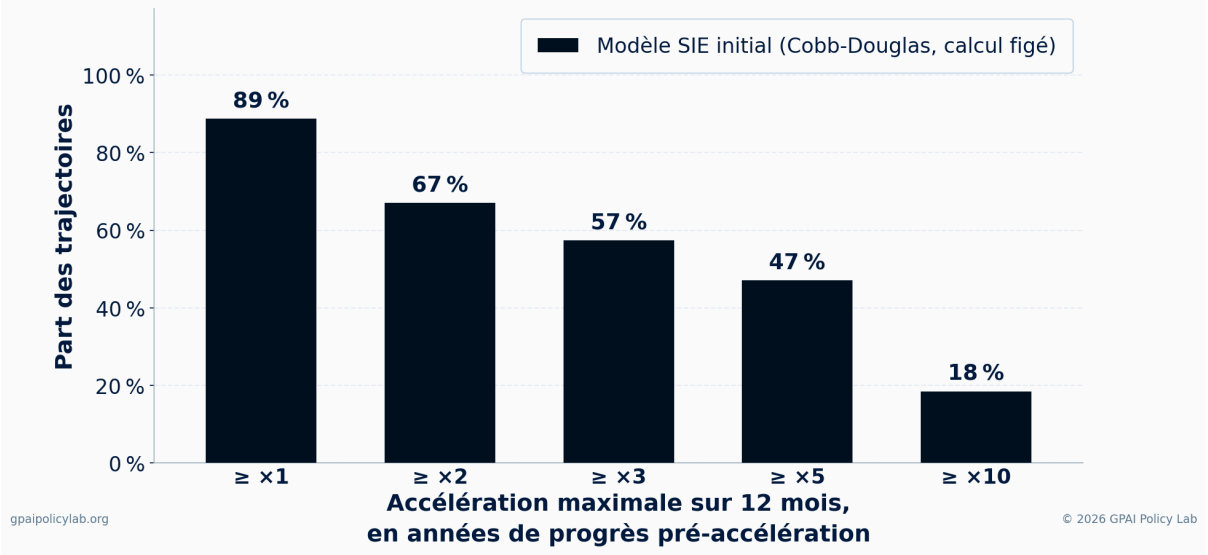

Figure 14 - Accélération maximale sur 1 an après automatisation totale de la R&D en IA dans le modèle SIE. Monte-Carlo sur 1000 trajectoires à partir des paramètres originaux du modèle via le code https://github.com/thoulden/Accelerated_AI_Progress

Le modèle SIE tend à considérer **qu'il est très probable (≈ 67%, soit deux chances sur trois) d'obtenir au moins un doublement de la vitesse de progression des capacités (c'est-à-dire qu'en une année on progresse autant qu'en deux années avant l'automatisation de la R&D).** Le modèle estime enfin à environ 1 chance sur 2 une accélération au moins **quintuplée** (>x5) des progrès (1 an de progrès en moins de 3 mois) et environ 1 chance sur 5 d'une **accélération décuplée** (>x10).

Précisons enfin que le modèle prévoit dans de rares cas (environ 1 chance sur 10) des scénarios où il ne se produit pas d'accélération du progrès dans les capacités, ce qui équivaut à un statu quo, voire un ralentissement par rapport au rythme historique.

L'analyse du modèle SIE souligne qu'il est très probable que l'automatisation totale de la R&D en IA produise une accélération marquée de l'amélioration des capacités des systèmes d'IA, pouvant aller d'un doublement du rythme de progrès historique à un décuplement dans les trajectoires les plus agressives. Dans cette perspective, le rythme actuel et passé de progrès, déjà rapide, n'est pas un bon indicateur du rythme futur et ne permet pas d'anticiper la rapidité de l'évolution des capacités des systèmes d'IA, et donc les risques émergents. Il existe toutefois des limites au modèle, notamment en ce qui concerne l'indépendance totale du progrès par rapport à la puissance de calcul (hypothèse de la substituabilité entre puissance de calcul et progrès algorithmique).

SIE modifié pour tenir compte de la puissance de calcul

La fonction de production dans SIE suppose une substituabilité unitaire (Cobb-Douglas) entre le travail et la puissance de calcul : l'effort de recherche R est $R(L, C) = L^{\alpha} C^{1-\alpha}$

Or, une partie significative des capacités de calcul est allouée par les entreprises à la frontière à la R&D en IA[87]. En effet, la découverte d'améliorations algorithmiques n'est pas indépendante de la puissance de calcul : au contraire, il est nécessaire de réaliser des expériences pour obtenir des résultats et explorer de nouvelles architectures, et ces expériences sont coûteuses en puissance de calcul[88]. On peut tenir compte de cette complémentarité en remplaçant la fonction Cobb-Douglas par une fonction CES qui introduit un facteur de complémentarité sigma qui détermine à quel point on peut remplacer de la puissance de calcul par simplement de meilleurs algorithmes (donc, plus concrètement du travail de chercheurs humains ou IA). Ce facteur de complémentarité vaut 1 par défaut dans le modèle de Davidson (cas Cobb-Douglas), et une valeur légèrement inférieure à 1 signifie que, si la puissance de calcul est figée, on ne peut pas entièrement compenser sa contribution au progrès des capacités par plus d'effort de recherche.

La fonction d'output recherche initiale était : $R(L,C) = L^{\alpha}C^{1-\alpha}$

La forme CES devient : $R(L,C) = [\alpha L^{\rho} + (1-\alpha)C^{\rho}]^{1/\rho}$ avec $\rho = \frac{\sigma-1}{\sigma}$

Sachant que $r := \frac{\alpha\lambda}{\beta}$ et que α correspond à l'élasticité unitaire (constante) de la fonction Cobb-Douglas, il suffit de remplacer α par l'élasticité variable de la fonction CES qu'on évalue en chaque combinaison (L,C) et qui dépend du ratio L/C. Comme C=1 (compute constant), on peut se concentrer sur L, qui double à chaque itération.

Cette élasticité vaut : $\theta(L) = \frac{\partial ln(R)}{\partial ln(L)} = \frac{\alpha L^{\rho}}{\alpha L^{\rho} + (1-\alpha)}$

Le nouveau paramètre est donc $r_{CES} = \frac{\lambda\theta(L)}{\beta}$ et donc $r_{CES} = r * \frac{\lambda\theta(L)}{\alpha}$

Le modèle se comporte donc comme le modèle Cobb-Douglas avec en plus une pénalité sur r car $\frac{\lambda\theta(L)}{\alpha} < 1$ dès que L augmente. Comme r est déjà décroissant par défaut dans le modèle de Davidson (décroît linéairement à mesure que l'on s'approche du plafond) la décroissance est encore accrue, ce qui explique que r passe très rapidement sous 1, provoquant l'arrêt de l'accélération de la croissance.

Étant donné que la puissance de calcul ne sera probablement pas constante, on doit ajouter pour plus de réalisme cette variable dans l'output du modèle (Cobb Douglas et CES) via Progrès(t) = software(t) + croissance de la puissance de calcul(t). Cette comptabilité permet une comparaison pertinente dans le mode CES, et peut également être utilisée pour le mode Cobb-Douglas. On transforme également le modèle SIE discrétisé en un modèle continu, ce qui permet d'éviter l'instabilité numérique lorsque L double très rapidement.

[87] Denain et Wu. "Final training runs account for a minority of R&D compute spending." Epoch AI, Gradient Updates (2026). https://epoch.ai/gradient-updates/r-and-d-vs-training-compute

[88] Gundlach et al. "On the Origin of Algorithmic Progress in AI." *arXiv* (2025). https://arxiv.org/abs/2511.21622

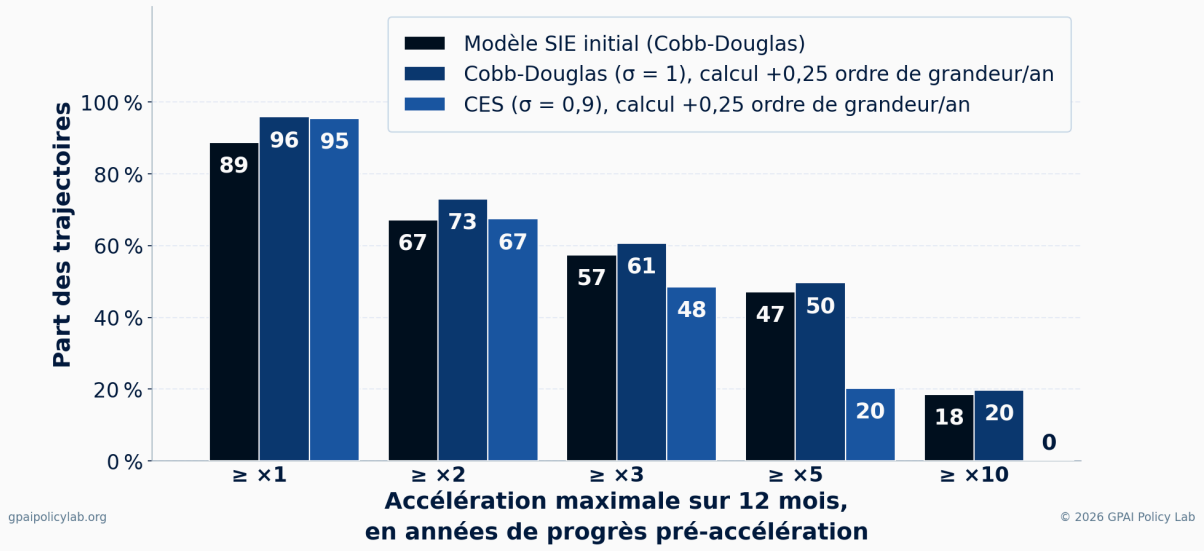

Figure 15 - Comparaison du modèle de SIE d'origine (évolution de la puissance de calcul non prise en compte) et du modèle modifié (complémentarité entre puissance de calcul et effort de recherche). On remarque qu'une valeur très proche mais inférieure à 1 (sigma = 0,90) suffit déjà à réduire très fortement la proportion de scénarios très rapides (de 50% à 20% de trajectoires avec accélération supérieure à x5, aucune trajectoire >x10).

Pour les scénarios à croissance de la puissance de calcul, on retient la valeur de 0,25 ordre de grandeur par an, soit un doublement tous les 14 mois environ (ce qui correspond à l'hypothèse du modèle AIF post-2028). Concernant le choix de sigma (facteur de complémentarité), une valeur plus basse contraindrait encore davantage la vitesse de croissance, mais il est difficile d'estimer précisément la valeur empirique en raison du nombre limité d'informations communiquées par les entreprises à la frontière sur leurs méthodes de recherche[89]. Cette démonstration sur la sensibilité de l'accélération à la puissance de calcul disponible permet d'établir deux constats : d'une part, qu'il s'agit d'un élément déterminant pour l'accélération rapide des capacités dans le paradigme actuel d'entraînement des modèles d'IA, d'autre part que même en tenant compte de cette dépendance on observe régulièrement un doublement (67% des trajectoires) voire un triplement (48% des trajectoires) du rythme d'évolution des capacités.

B. Détails du modèle AI Futures (AIF)

Le modèle SIE a l'avantage d'être un modèle fortement simplifié, mais il ne décrit pas de manière fine les mécanismes qui contribuent à l'automatisation de la R&D. Par ailleurs, il manque de granularité dans sa description des capacités : **exprimer un progrès futur en termes d'années au rythme pré-automatisation ne permet pas vraiment de se rendre compte des niveaux de capacités atteints et de leur signification par rapport aux seuils de risque**. Enfin, le modèle SIE simplifie fortement le processus d'automatisation de la R&D en ne tenant pas compte de la complémentarité entre le travail humain et le travail réalisé par IA dans la conception des

[89] On peut toutefois supposer que sigma << 1 sur la base de Whitfill et Wu, op. cit.

expériences, surtout lors de la phase où les capacités humaines sont encore indispensables aux percées théoriques. Le modèle AI Futures (AIF), plus complexe, développe une approche plus réaliste pour capturer les déterminants de l'automatisation de la R&D.

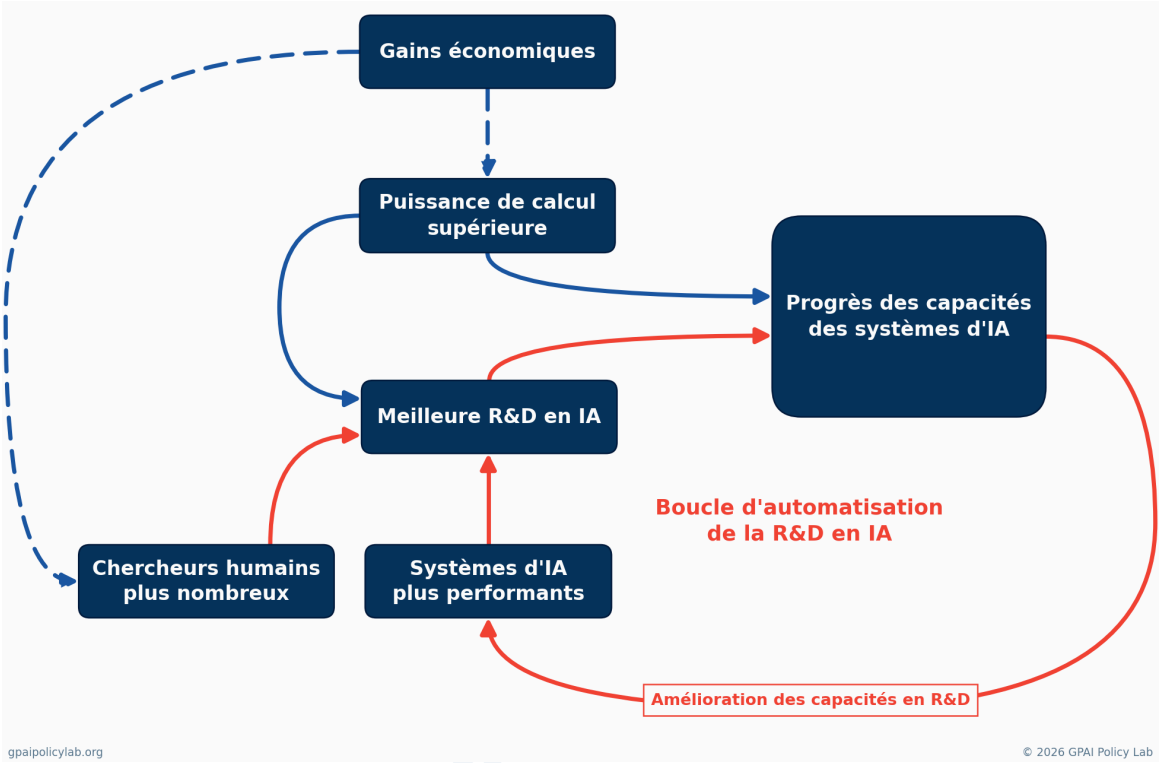

Figure 16 - Schéma simplifié du modèle AI Futures (AIF). La boucle de rétroaction principale reste la boucle d'automatisation de la R&D en IA (en rouge), mais le cadre de modélisation tient également compte de la croissance de la puissance de calcul disponible et du nombre de chercheurs humains, également essentiels à l'amélioration des capacités de l'IA. Celle-ci est modélisée par des séries exogènes (calculées a priori, sans boucle de rétroaction économique), représentées en pointillés.

La boucle principale par laquelle les capacités augmentent est donc celle de l'automatisation de la R&D en IA (en rouge), mais les contraintes de puissance de calcul et de ressources humaines sont également prises en compte, ce qui contribue au réalisme du modèle. Dans le modèle AIF, les flux qui correspondent à la boucle de rétroaction économique sont exogènes, et calculés a priori via des séries temporelles qui se basent sur les hypothèses des auteurs. Ainsi, les flux de puissance de calcul croissent à environ 0,6 ordres de grandeur (OOM) par an jusqu'à 2028 (taux historique), puis à 0,25 OOM/an ensuite. Les flux correspondant à la population de chercheurs humains sont également exogènes. Pour un détail précis de l'architecture du modèle et les valeurs choisies pour les séries temporelles, voir :

https://docs.google.com/document/d/1ru6Okbxb6XuH18Cz8439sdQJazMV39hNxsWDokh97r0/edit

La phase d'accélération rapide des progrès est jalonnée par plusieurs bornes qui constituent des **seuils de capacités de l'IA qu'on peut mettre en équivalence avec les niveaux d'agentivité proposés dans le Tableau 2**. Parmi ces bornes, trois constituent des étapes fondamentales :

- **La première phase commence à l'initialisation du modèle (2017) et s'achève par l'automatisation totale du code informatique. Cette borne, appelée « Codeur Automatisé » est définie comme un système d'IA en capacité d'automatiser de manière autonome toutes les tâches de code réalisables par des ingénieurs en informatique**, quelle que soit leur durée et leur complexité. Cette borne correspond environ au niveau d'autonomie A4 (*agent orchestrateur*) dans l'échelle élaborée dans cette note (Tableau 2) et est définie par rapport aux horizons temporels METR à 80% de succès, en supposant que la croissance est presque exclusivement super-exponentielle (c'est à dire que chaque doublement d'horizon prend un temps plus faible que le précédent).
- La seconde phase étudie l'automatisation totale de la R&D en IA, à partir du moment où le code informatique est automatisé. **Elle s'achève par la borne "Chercheur IA surhumain", qui qualifie un système dont les capacités en recherche sont supérieures à celles du meilleur chercheur en IA**. Une telle performance dans le domaine de la R&D en IA se situe probablement à la frontière entre le niveau A4 (*agent orchestrateur)* et A5 (*agent autonome - AGI*), **mais ne franchit pas le seuil AGI-SC en raison d'une spécialisation encore très importante et donc de performances plus faibles hors de ce domaine[90].**
- La troisième phase correspond à l'accroissement rapide des capacités. **Elle contient la borne TED-AI (Top expert dominating AI) dont la définition correspond à AGI-SC : il s'agit d'un système d'IA dont le niveau sur toutes les tâches cognitives est au-dessus du meilleur professionnel dans chaque domaine**. Un tel système est donc théoriquement en mesure d'automatiser l'entièreté du travail cognitif. Cette phase se poursuit jusqu'à l'atteinte des limites absolues de performance (bien au-delà de la borne AGI-SC, ralentissement puis stagnation du progrès).

Le modèle AI Futures introduit un mécanisme qui décrit l'évolution de l'intuition en recherche (*research taste*), afin de mieux rendre compte de la différence entre l'automatisation du code (déjà en cours, phase 1) et celle de la recherche (plus complexe, phases 2 et 3). Les modèles actuels présentent déjà des capacités de programmation exceptionnelles. On pourrait s'attendre à ce que ces performances suffisent à automatiser entièrement la R&D en IA, mais ce n'est pas le cas. En effet, les chercheurs humains disposent à ce stade d'un avantage significatif sur les systèmes d'IA actuels, qui réside en partie dans leur intuition. Bien que nettement moins rapides et moins capables d'accumuler un grand nombre d'informations, les chercheurs humains ont, par leur expérience, développé une capacité à évaluer très en amont le caractère prometteur ou non d'une idée, sans avoir à l'explorer en profondeur. Cette capacité manque encore aux meilleurs systèmes d'IA et constitue un obstacle à l'automatisation totale de la recherche, surtout pour résoudre des problèmes « ouverts » et difficiles à vérifier rapidement. Pour autant, les systèmes d'IA les plus avancés

[90] L'évaluation du niveau d'autonomie d'un système à la borne "Chercheur surhumain" dans AIF est difficile, et dépend essentiellement du caractère plus ou moins irrégulier du profil de performance du système d'IA (soit encore très spécialisé (jagged), soit au contraire polyvalent et homogène). Le caractère encore très hétérogène des modèles actuels tend à faire penser qu'un système capable d'automatiser la R&D en IA n'est pas encore au-dessus du niveau humain dans l'entièreté des domaines cognitifs.

paraissent progresser dans cette direction. **Il est donc essentiel de modéliser la progression de l'intuition pour la recherche si l'on veut tenir compte de cette asymétrie entre chercheurs humains et systèmes d'IA**. C'est ce que vise le modèle AI Futures : une partie significative du travail de modélisation vise à estimer à quelle vitesse progresse l'intuition des systèmes d'IA spécialement conçus pour la R&D, et à quelle intensité les boucles de rétroaction se mettent en place (plus le système possède une bonne intuition, plus il est en mesure de développer une génération de systèmes encore meilleurs, et ainsi de suite). Cette modélisation est hautement spéculative en raison de la forte confidentialité concernant les capacités en R&D des meilleurs systèmes. Les auteurs du modèle s'appuient sur un grand nombre d'indicateurs de performance dans des domaines divers et observent leur progression historique, puis extrapolent la vitesse de progression qu'on peut raisonnablement attendre dans le domaine de l'intuition pour la R&D en IA[91].

L'unité de mesure du paramètre "research taste" est construite comme suit : on considère les 1000 meilleurs chercheurs d'une entreprise d'IA et l'on suppose que leur capacité à extraire de l'information d'une expérience donnée suit une distribution log-normale. Le meilleur chercheur est donc top 1/1000, soit 3,09 SD sur la loi normale sous-jacente. On définit l'écart entre le top 1/1000 et le chercheur médian (0 SD, 500/1000) comme un "écart meilleur-médian". L'écart meilleur-médian est un paramètre libre estimé entre x3 et x4 par les auteurs du modèle, et extrapolé par la suite au-delà de l'intervalle humain via l'identité 3,09 SD = un écart médian-meilleur.[92]

Une manière de représenter l'incertitude inhérente à la paramétrisation du modèle est de tracer un grand nombre de trajectoires dont les paramètres initiaux ont été choisis dans un intervalle jugé plausible. Cette méthode permet d'obtenir l'ensemble de trajectoires suivantes, via la paramétrisation Eli 04/26, qui correspond à la paramétrisation la plus conservatrice du modèle.

[91] AI Futures Project. "[AI Futures Model] Supplementary materials." (2025), p. 47-54. https://docs.google.com/document/d/1ru6Okbxb6XuH18Cz8439sdQJazMV39hNxsWDokh97r0/edit

[92] On peut prolonger artificiellement la métrique "performance" en utilisant la fonction CDF

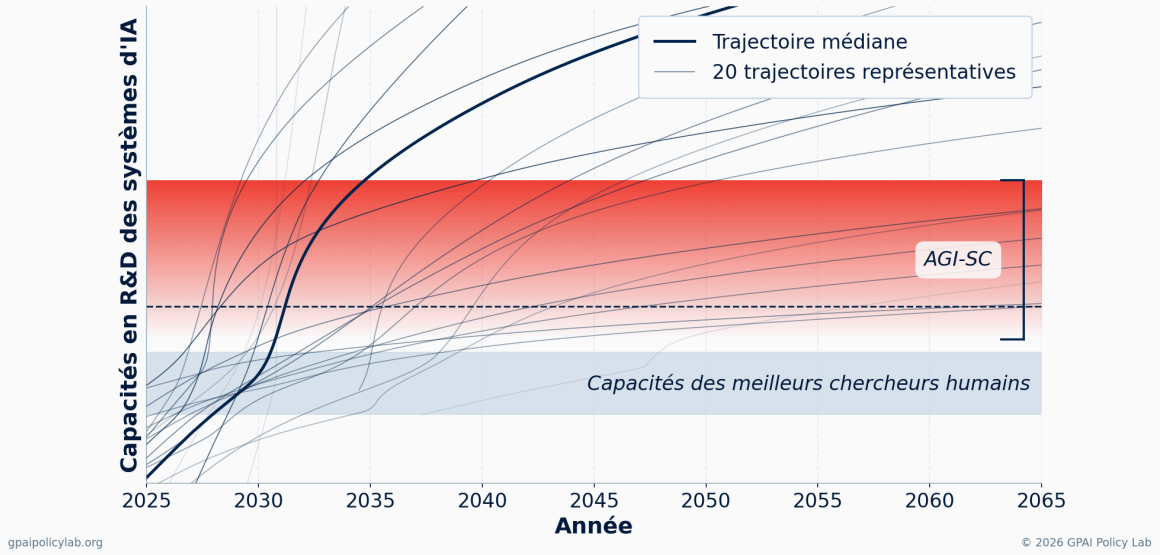

Figure 17 - Ensemble de trajectoires du modèle AIF générées à partir de la distribution de paramètres d'Eli Lifland (avril 2026). La trajectoire centrale correspond à la médiane tous paramètres du Monte-Carlo (p50 pour chaque variable). La bande "capacités humaines" figure entre -3,1 et 3,1 SD, ce qui correspond à la distribution de la population théorique de chercheurs par rapport à laquelle le research_taste est défini (rang 1000/1000 à rang 1/1000). La variable tracée est la valeur en déviations standard par rapport à la population de référence des capacités en R&D en IA du meilleur système disponible au temps t. La bande rouge AGI-SC est dérivée à partir de la borne TED-AI du modèle, et des paramètres choisis dans le Monte-Carlo.

Sur cette agrégation de simulations, on observe clairement la large dispersion des trajectoires. Faire figurer plusieurs trajectoires individuelles permet de visualiser les différentes accélérations de performance qui peuvent être atteintes : certaines marquent une forte accélération (quasi-verticales, franchissement du seuil entre 2027 et 2030) tandis que d'autres ont une croissance régulière (rythme constant) voire ralentie. Notons que dans la trajectoire médiane, les capacités en R&D des meilleurs chercheurs en IA sont dépassées avant 2031, et le seuil AGI-SC quelques mois après.

L'effort de réalisme du modèle AI Futures permet de souligner la complexité du processus d'automatisation totale de la R&D dans lequel de nombreuses frictions peuvent intervenir (complémentarité recherche et puissance de calcul, progression incertaine des capacités de l'IA en R&D, automatisation progressive de la programmation informatique…). Pour autant, plus de la moitié des scénarios franchissent la catégorie AGI-SC avant 2035, en raison de la forte boucle de rétroaction qui se met en place une fois que les capacités **des systèmes d'IA en R&D dépassent de loin celles des meilleurs chercheurs.** Cette accélération rapide post-automatisation concorde avec les trajectoires du modèle simplifié SIE et nécessite une vigilance particulière en raison de la forte discontinuité avec la progression des capacités au rythme historique.

C. Détails du modèle GATE

Les modèles précédents (SIE et AIF) ne tiennent pas compte des flux économiques ; or, la quantité d'investissements nécessaire pour automatiser l'entièreté de la R&D en IA pourrait être supérieure aux capacités de financement des entreprises développant les IA de frontière, qui malgré leurs revenus exceptionnels sont actuellement déficitaires[93]. Le souci de réalisme économique est l'objet du modèle GATE (*Growth and AI Transition Endogenous model*), qui combine deux approches pour prédire l'évolution des capacités de l'IA : **d'une part une approche basée sur la progression des capacités** (proche de celle du modèle AI Futures), **d'autre part une structure macro-économique qui simule les flux financiers entrants (investissement) et sortants dans l'économie mondiale, à mesure qu'un nombre croissant de tâches est automatisé par des systèmes d'IA. Le cœur du modèle est un planificateur social qui alloue la production totale entre investissement dans l'IA (puissance de calcul, R&D en software, R&D hardware), investissement en capital conventionnel et consommation en fonction des rendements anticipés sur chaque type d'investissement**[94].

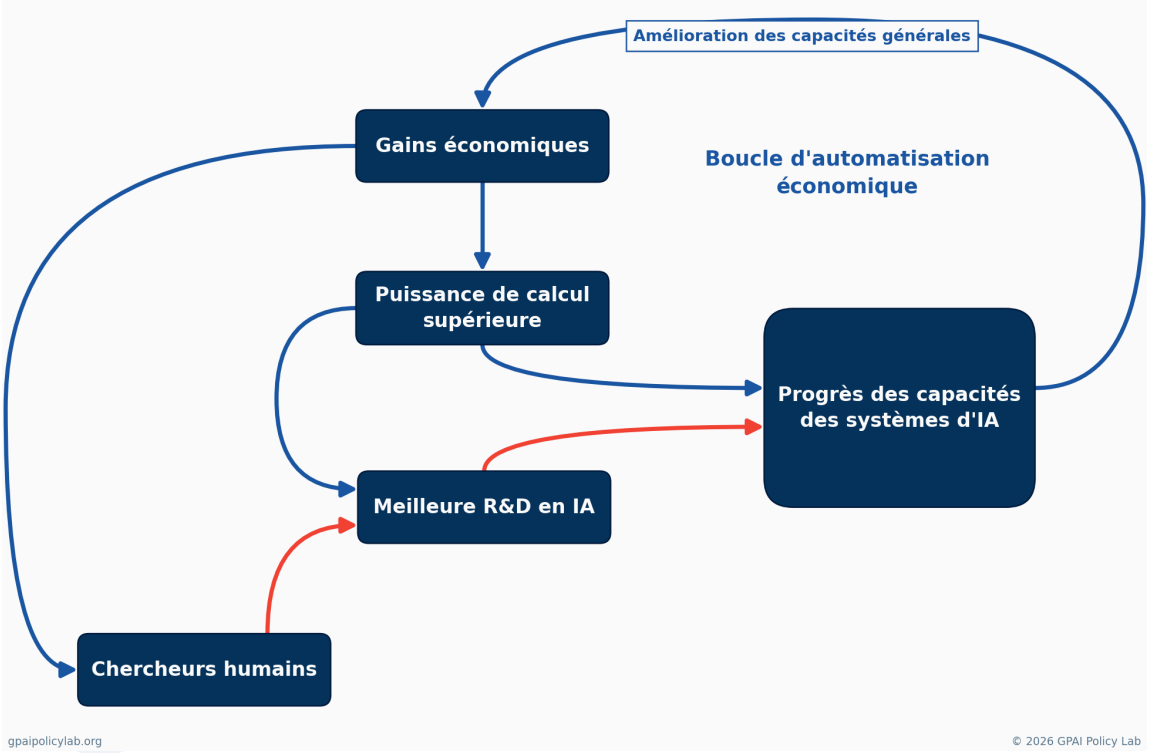

Figure 18 - Schéma simplifié du modèle GATE. La boucle de rétroaction principale est ici, contrairement aux autres modèles, la boucle économique : à mesure que les capacités progressent, des tâches plus complexes sont automatisées dans l'économie, ce qui provoque une forte croissance de la production, réinvestie ensuite dans les infrastructures de calcul et dans la R&D, mais uniquement via des facteurs humains (donc pas de boucle de rétroaction d'automatisation de la R&D en IA par l'IA).

[93] Edwards. "OpenAI's financials have leaked, showing $21 billion in losses against $13 billion in revenue." Fortune (2026). https://fortune.com/2026/06/16/openai-financials-leaked-losses-revenue-profit/

[94] Erdil et al. "GATE: An Integrated Assessment Model for AI Automation." arXiv (2025). https://doi.org/10.48550/arXiv.2503.04941

GATE modélise également les frictions susceptibles de ralentir l'automatisation totale de l'économie, en particulier **les effets Baumol : lorsque seule une partie des tâches est automatisée, les tâches non automatisées tendent à devenir les facteurs limitants qui plafonnent la croissance globale**. Ce mécanisme constitue l'objection économique standard aux scénarios d'accélération rapide. **Or les simulations de GATE montrent que ces effets sont plus faibles qu'anticipé**[95]. L'ajout de frictions supplémentaires (externalités de R&D, incertitude des investisseurs, coûts de réallocation du travail) atténue l'ampleur de l'accélération économique mais ne la supprime pas.

Le résultat le plus frappant de GATE concerne le niveau d'investissement économiquement justifié. Compte tenu des flux de salaires versés aux travailleurs dans l'économie mondiale, dont une fraction substantielle pourrait être captée par l'automatisation (jusqu'à 100%), les simulations de référence[96] suggèrent q**u'un investissement optimal annuel dans l'IA pour 2026 se situerait entre 10 et 30 fois au-dessus du niveau d'investissement actuel.**

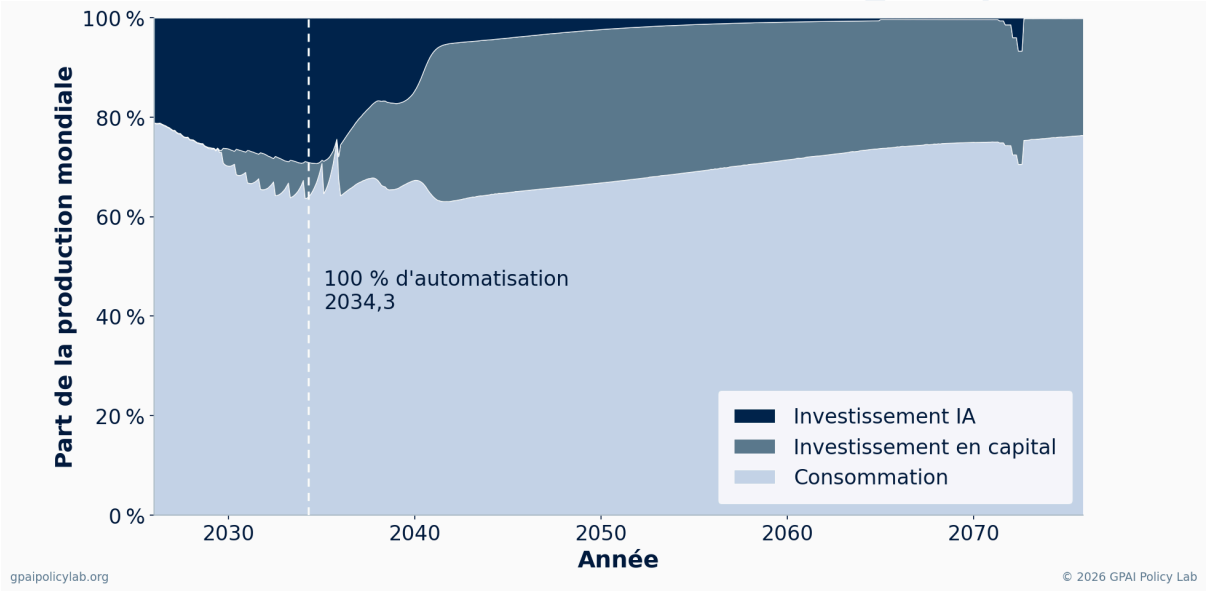

Figure 19 - Allocation optimale dans le scénario tous paramètres médians du modèle GATE. La quantité d'investissement dans l'IA jugée optimale par le planificateur social correspond à 20% de la production mondiale dès l'année d'initialisation (2026). Cette valeur irréaliste (l'investissement actuel s'élève à environ 1-2% de la production mondiale totale) s'explique par le caractère omniscient du planificateur qui anticipe très tôt les rendements associés à l'automatisation totale du travail, et priorise donc un transfert d'investissement dans le domaine de l'IA (disparition totale de l'investissement conventionnel).

Cette divergence d'investissement entre le modèle GATE et ce qui est observé peut s'expliquer par plusieurs facteurs (pas d'investissement mondial centralisé, anticipation de barrières réglementaires, incertitude sur le degré d'automatisation possible des tâches physiques, aversion au risque des investisseurs), mais elle conduit à un renversement de perspective : le niveau d'investissement actuel apparaît vraisemblablement plutôt en sous-régime qu'en sur-régime. **Si tel est le cas, l'hypothèse**

[95] Potlogea et Ho. "AI and explosive growth redux." Epoch AI, Gradient Updates (2025). https://epoch.ai/gradient-updates/ai-and-explosive-growth-redux
[96] Élaborées par les auteurs du modèle (Epoch AI), voir https://epoch.ai/gate

d'une « bulle IA » souvent évoquée s'éloignerait au profit de celle d'un rattrapage progressif vers un niveau d'investissement économiquement justifié par le gisement de valeur associé à l'automatisation du travail.

On peut également faire l'hypothèse que le faible niveau d'investissement actuel (relativement aux modélisations) est dû à une forte incertitude sur les capacités à automatiser effectivement une grande partie du travail humain. Cette possibilité peut être intégrée dans GATE en attribuant au planificateur social des croyances sur le niveau maximal d'automatisation atteignable. Par défaut, le planificateur attribue 100% de probabilité à une automatisation totale, mais l'on peut créer artificiellement des profils nettement plus conservateurs pour observer les effets de ces croyances sur l'allocation de la production. On peut ainsi établir un scénario identique au scénario médian mais en définissant une croyance de 95% de probabilité d'automatisation plafonnée à 20% des tâches de l'économie et seulement 5% au-delà. Le planificateur investit donc plus prudemment initialement, ne s'attendant qu'à un retour sur investissement limité par l'impossibilité d'automatiser au-delà de 20% de l'économie, puis met à jour ses croyances à mesure que le niveau d'automatisation dépasse le seuil de 20%. Même en tenant compte d'un tel profil de croyance, la variation d'investissement dans l'IA est très faible et simplement réduite lors des premières années, avant de rejoindre un niveau d'environ 20% dès que le planificateur social a perçu la possibilité d'automatiser au-delà de ses croyances initiales.

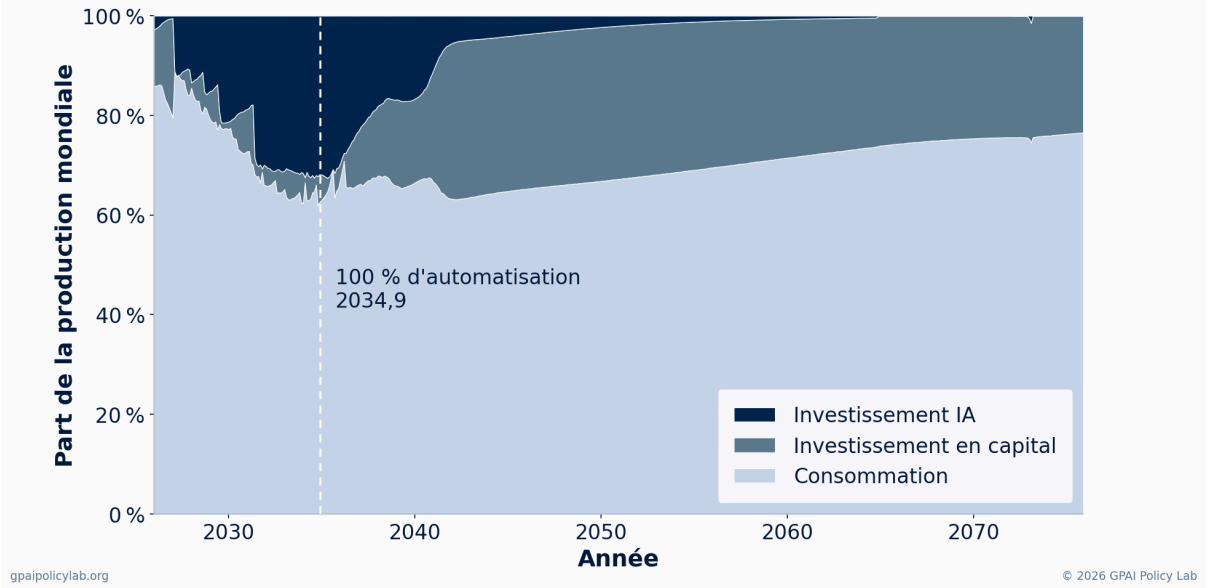

Figure 20 - Allocation optimale dans le scénario conservateur du modèle GATE. La quantité d'investissement dans l'IA à l'initialisation est nettement plus faible (≈3% contre 20% dans le scénario par défaut). Pourtant, l'investissement jugé optimal converge rapidement vers les niveaux observés dans le scénario par défaut, à mesure que le planificateur social modifie ses croyances pour tenir compte d'une automatisation plus grande qu'anticipée. Malgré des croyances très conservatrices et un investissement plus faible pendant les 2-3 premières années, la date d'automatisation totale ne se décale que de quelques mois.

GATE joue un rôle complémentaire par rapport aux modèles SIE et AIF, en prenant en compte les contraintes macro-économiques à l'investissement qui sont nécessaires à l'augmentation des capacités. La modélisation GATE met en évidence que les trajectoires de capacités décrites par SIE et AIF sont atteignables économiquement. Autrement dit, la dynamique d'accélération des investissements en IA ne bute pas sur une contrainte macroéconomique qui viendrait naturellement la réguler : les flux de capitaux et le déploiement d'infrastructures de calcul associés à l'automatisation totale du travail restent cohérents avec la logique économique d'allocation des investissements. GATE apporte des arguments solides en défaveur de la thèse du ralentissement des capacités en raison des freins économiques, et semble au contraire pointer vers une accélération de l'investissement dans les infrastructures, qui peut déjà être observée[97]. Cependant, il est essentiel de préciser que GATE considère le potentiel d'automatisation plutôt que l'automatisation effective : le modèle simplifie fortement les frictions au déploiement dans l'économie réelle de systèmes pleinement autonomes, et considère que dès lors qu'un type de tâche économiquement viable est automatisable, il le devient immédiatement. Pour cette raison, le pourcentage d'automatisation dans GATE devrait être interprété comme un seuil de capacités plutôt que comme un seuil d'automatisation.

Les trajectoires de GATE étayent les conclusions du modèle AIF, en convergeant vers l'émergence probable de systèmes d'IA capables d'automatisation totale de l'économie mondiale avant 2040 (environ 70% des trajectoires), et ce malgré l'effort de réalisme en termes de contraintes économiques. GATE rend explicites les mécanismes économiques qui justifient l'ampleur des investissements dans l'infrastructure IA : le revenu associé au travail humain est tel qu'en capter un pourcentage même faible initialement permet de générer un profit suffisant pour investir massivement dans des infrastructures extrêmement coûteuses qui seront amorties sur la durée. Si ces prédictions peuvent sembler particulièrement agressives, les trajectoires actuelles du profit généré par les entreprises à la frontière sont pourtant alignées sur la dynamique projetée par GATE.

Enfin, GATE tend à confirmer l'importance de la puissance de calcul pour l'automatisation totale des tâches à valeur économique, d'une part pour ce qui est de l'entraînement de systèmes d'IA plus performants, d'autre part pour l'inférence, c'est à dire la capacité à faire fonctionner un grand nombre d'instances d'un système d'IA sur des tâches longues et exigeantes. En effet, même après avoir entraîné des systèmes aux performances globalement surhumaines, il est nécessaire de s'assurer d'une puissance de calcul suffisante pour automatiser effectivement toute l'économie. Ce constat complémentaire confirme le caractère hautement critique de la puissance de calcul et de sa répartition entre États.

[97] Juniewicz. "Hyperscaler capex is on trend to outpace their cash inflows by the end of 2026." Epoch AI, Data Insights (2026). https://epoch.ai/data-insights/hyperscaler-capex-vs-cash-flow

D. Harmonisation des trois modèles

Les trois modèles étudiés traitent chacun d'un aspect spécifique des processus d'automatisation de la R&D en IA, ce qui les rend difficiles à comparer directement.

- **Le modèle SIE cherche à modéliser le plus simplement possible la dynamique d'accélération une fois l'automatisation de la R&D en IA achevée.** Il génère uniquement des **trajectoires d'accélération**, qui indiquent combien d'années au rythme de progrès historiques ont été réalisées en une seule année après l'automatisation totale.
- **Le modèle AIF développe une mécanique complexe pour anticiper l'évolution de l'intuition pour la recherche (paramètre déterminant pour l'automatisation totale de la R&D)** et reproduit les trajectoires de capacités à partir de 2017 pour tenir compte des progrès ayant déjà eu lieu et de la **complémentarité entre puissance de calcul et automatisation de la programmation**. Il contient une borne correspondant au seuil **AGI-SC**, définie à partir **d'un seuil de capacité uniquement exprimé en termes de performance en R&D en IA.**
- Le modèle GATE s'appuie sur des mécanismes macro-économiques et génère des trajectoires d'automatisation allant jusqu'à l'automatisation totale du travail humain (cognitif et physique), à partir d'une hypothèse forte de planification économique centralisée, sans modéliser les boucles de rétroaction de l'automatisation de la R&D en IA (seuls les investissements peuvent augmenter les progrès des capacités, on ne tient pas compte de la contribution des systèmes d'IA déjà existants dans le domaine de la R&D en IA).

Pour chaque modèle, **la représentation en boîtes à moustaches (Figure 7) se fait sur l'ensemble des 1000 trajectoires et non sur les seules trajectoires qui atteignent le seuil AGI-SC avant une date donnée**. Ainsi le dernier décile ne figure pas pour les modèles SIE et AIF, n'étant pas atteint dans la plage de temps fixée (une partie des scénarios n'atteint jamais la borne, quel que soit l'horizon temporel).

- Pour AIF, la paramétrisation utilise les 1000 trajectoires de la paramétrisation d'Eli Lifland (avril 2026) en prenant pour date AGI-SC le franchissement de la borne TED-AI par détection via la métrique "intuition pour la recherche" (*research taste*).
- Pour GATE, on conditionne sur 100% d'automatisation, en prenant en compte les limites de réalisme du modèle et l'absence de delta entre capacités et déploiement puis rendements dans l'économie. C'est un seuil qui semble au premier abord nettement plus exigeant que l'automatisation totale du travail cognitif (seuil AGI-SC)[98]. Cependant, la complexité des tâches dans GATE est décrite par leur intensité en puissance de calcul nécessaire pour les compléter. Cela amène à supposer qu'il est possible que les tâches intellectuelles les plus difficiles soient au moins aussi dures que les tâches manuelles les plus difficiles (par exemple, R&D dans les technologies de pointe, découvertes scientifiques, etc.). Si cette hypothèse est juste, alors c'est le seuil de 100% des tâches qui doit être considéré comme la borne pour l'automatisation cognitive totale (AGI-SC), en interprétant ce seuil comme un seuil de capacité mais pas de déploiement effectif dans l'économie. La paramétrisation est réalisée en conservant toutes les variables telles que définies par les auteurs du modèle, et

[98] Il semble probable que les tâches cognitives à valeur économique ne représentent qu'environ 50-70% de la masse totale des tâches salariées en volume financier, et que les tâches restantes sont de nature physique / hybride, ce qui est plus lent à automatiser (nécessite le développement d'infrastructures robotisées).

en mettant à jour les conditions économiques initiales à partir des données sur l'année 2026.

- Pour SIE, on procède comme suit :
  - on part de la paramétrisation initiale de Davidson
  - on détermine la date calendaire d'automatisation totale de la R&D en IA via le modèle AIF, qui contient déjà une borne équivalente (intitulée "Chercheur IA surhumain" qu'on note ici par l'acronyme "SAR" pour *Superhuman AI researcher*). On prend l'ensemble des dates de la borne SAR dans la configuration AIF (paramétrisation Eli) et on en déduit une distribution de dates plausibles pour t0 SIE (ASARA, qui correspond à SAR).
  - détermination de l'écart entre t0 SIE - AGI-SC via AIF : on regarde l'écart en métrique effective flops entre la borne SAR et la borne TED-AI qu'on normalise ensuite par rythme moyen en eFLOPs/an (sur l'année 2024-2025) pour chaque trajectoire. On obtient une distribution du "nombre d'années au rythme historique" nécessaire pour franchir l'écart entre la phase "automatisation totale de la R&D" et la phase "profil de compétence supérieur aux experts dans tous les domaines cognitifs", c'est-à-dire AGI-SC.
  - On utilise ensuite l'output classique de SIE en années au rythme historique pour obtenir la date calendaire de franchissement du seuil AGI-SC.

La méthode utilisée pour SIE reste fondamentalement très dépendante des résultats du modèle AIF, mais **cela tient à la difficulté de transcrire un facteur d'accélération en date calendaire**. Par ailleurs, les trois modèles partagent des hypothèses fortes (croissance des capacités comme le logarithme de la puissance de calcul effective, améliorations algorithmiques importantes possibles, possibilité réelle d'automatisation de tout le travail humain), qui sont déterminantes dans les trajectoires obtenues. Pour autant, ces hypothèses de travail apparaissent crédibles au regard des progrès historiques des systèmes d'IA.

E. Classification des systèmes d'IA par niveau d'agentivité

La complexité de la construction d'une échelle d'agentivité correcte réside dans la nécessité de tenir compte des performances surhumaines des meilleurs systèmes d'IA actuels dans des domaines où le feedback est dense et/ou vérifiable (par exemple code informatique) mais de performances limitées dans des domaines qui nécessitent une planification long terme, une gestion de l'incertitude et des comportements exploratoires[99]. L'échelle présentée dans le tableau 2 ne dispose pas en l'état de métriques permettant de situer exactement un système d'IA sur un niveau d'agentivité. Plusieurs métriques alternatives pourraient remplir cet usage, par exemple l'**Epoch Capabilities Index** ou les **horizons temporels METR**, mais leur structure continue ne permet pas de distinguer les ruptures entre chaque échelon. Pour préciser la nature de ces ruptures, on peut reformuler chaque échelon sur le principe de la capacité d'un système d'IA à fermer des boucles de correction entre son modèle de l'environnement et l'environnement dans lequel il évolue, afin d'atteindre un objectif de haut niveau pouvant être sous-spécifié (en général, un objectif mal spécifié est plus difficile à atteindre et nécessite donc des capacités de planification et d'exploration supérieures). Ainsi, le niveau

[99] Ou & Burnham. "AI doesn't get better at this board game with practice." Epoch AI (2026). https://epoch.ai/publications/earthborne-rangers-benchmark

d'agentivité d'un système détermine la complexité maximale de l'environnement dans lequel il peut atteindre un objectif donné.

On peut explorer l'hypothèse suivante pour la transcription des niveaux de l'échelle d'agentivité :

NIVEAU DE L'ÉCHELLE	TYPE DE BOUCLE DE RÉTROACTION	MÉCANISME SOUS-JACENT	BENCHMARKS REPRÉSENTATIFS
Systèmes non agentiques			
A0. IA-Outil (système spécialisé)	Génération ou décision réactive à partir d'une entrée ; le signal de récompense est dense et immédiat. Inclut les systèmes spécialisés surhumains, dont l'environnement est entièrement spécifié.	Apprentissage supervisé, apprentissage par renforcement en environnement fermé	Jeux à règles fixes
Systèmes faiblement agentiques			
A1. Système conversationnel	Génération sans auto-évaluation ; le système tient le contexte (sous condition de volume) et peut éventuellement demander des précisions, mais la correction vient de l'interlocuteur.	RLHF, Fine-tuning	Tests de connaissance sans raisonnement approfondi (MMLU, Winogrande, PIQA)
A2. Agent minimal	Boucle interne intra-épisode. Auto-évaluation, retour en arrière, exploration de l'espace des solutions ; Permet une planification courte sur tâche bien définie.	Test-time compute, Chain-of-thought	Tests de connaissance avec raisonnement (GPQA Diamond, VPCT) Tests de raisonnement abstrait (ARC-AGI 1, ARC-AGI 2)
Systèmes agentiques limités			
A3. Agent planificateur	Boucle environnementale intra-tâche. Correction sur un signal externe obtenu par usage d'outils (test qui échoue, commande qui renvoie une erreur).	RL sur tâches longues avec outils et harnais,	Time-Horizon METR, Mirror Code, SWE-Bench, GSOBench, Frontier Maths 1-4
A4. Agent orchestrateur	Boucle stratégique intra-mission. Re-décomposition du problème, révision des sous-objectifs et non seulement des actions ; exploration d'un environnement dont la structure est à découvrir ; gestion de l'incertitude ; orchestration de sous-agents.	Construction d'un modèle du monde explicité, entraînement multimodal, coordination multi-agents	ARC-AGI3, MazeBench2026, Vending Bench2, RLI, OpenProblems
Systèmes AGI au seuil critique (AGI-SC)			
A5. Agent autonome (AGI)	Boucle trans-épisode. Mémoire exploitée pour la correction ; horizon de planification excédant la fenêtre effective de traitement malgré un signal sparse au niveau de l'épisode ; gestion des échecs et exploration ciblée quand	Apprentissage continu ?	EBR Bench, Project Vend, Alpha Arena

	chaque essai nécessite d'allouer des ressources.		
A6. Agent autonome surhumain (AGI Forte)	Boucle d'acquisition. Identifie ses propres absences de compétence et les comble activement ; métacognition et compréhension de ses incertitudes ; maîtrise d'environnements complexes et adversariaux peuplés d'une multitude d'agents, avec coordination à plusieurs niveaux.	Mécanismes inconnus	Objectifs directement testés en conditions réelles ?

La formalisation d'une telle échelle est encore à l'état exploratoire, étant donné la difficulté de mesurer spécifiquement le degré d'agentivité d'un système en dehors d'une approche empirique basée sur les performances en environnement de test. En pratique, la borne AGI-SC (maîtrise au niveau expert de tous les domaines cognitifs) semble requérir au moins un niveau A5, mais sur de nombreuses tâches spécifiques un niveau A4 voire A3 pourrait suffire pour une automatisation avancée.