THE ANTHROPIC INSTITUTE · ESSAI

Quand l'IA se construit elle-même

Traduction française d'après « When AI builds itself », publié par l'Anthropic Institute.

Pendant la majeure partie de l'histoire de l'IA, les humains ont piloté chaque étape de son développement. Mais chez Anthropic, nous déléguons une part croissante du développement de l'IA aux systèmes d'IA eux-mêmes, ce qui accélère notre travail.

Poussée assez loin, et à condition de disposer d'assez de puissance de calcul, cette tendance pointe vers un système d'IA capable de concevoir et de développer son successeur de façon entièrement autonome. C'est ce qu'on appelle l'*auto-amélioration récursive*. Nous n'en sommes pas là, et l'auto-amélioration récursive n'a rien d'inéluctable. Mais elle pourrait arriver plus tôt que la plupart des institutions ne s'y préparent.

En nous appuyant sur des benchmarks publics et sur des données internes à Anthropic jusqu'ici inédites, l'Anthropic Institute montre que l'IA accélère déjà le développement des systèmes d'IA. Un seul exemple : aujourd'hui, les ingénieurs d'Anthropic livrent en moyenne **8 fois plus de code par trimestre** qu'entre 2021 et 2025.

Les tendances techniques décrites ici suggèrent que les systèmes d'IA vont devenir beaucoup plus capables dans les années à venir. Ces tendances ont d'immenses implications. Une IA capable de se construire elle-même serait une étape majeure dans l'histoire des technologies — une étape qui pourrait apporter d'énormes bienfaits au monde, dans la

science, la santé et au-delà. Mais une auto-amélioration récursive complète pourrait aussi accroître le risque que les humains perdent le contrôle des systèmes d'IA. Si ces systèmes sont capables de construire entièrement leurs successeurs, la manière dont nous les sécurisons, les surveillons et façonnons leur comportement n'en devient que plus cruciale.

2021–2023

Construire le premier Claude

Aux débuts, le travail chez Anthropic ressemblait à celui de n'importe quelle entreprise tech : des gens écrivant du code et de la documentation sur leur ordinateur portable.

2023–2025

Les chatbots

Les premiers chatbots aidaient sur certaines parties du processus, comme générer de courts extraits de code que l'on copiait ensuite dans son éditeur de texte.

2025–2026

Les agents de code

À mesure que les agents gagnaient en capacité, ils ont pu écrire et modifier du code par eux-mêmes, parfois des fichiers entiers.

Aujourd'hui

Les agents autonomes

Les agents peuvent désormais exécuter eux-mêmes du code et déléguer des heures de travail à d'autres agents.

20XX ?

Boucler la boucle

À l'avenir, les agents pourraient devenir assez capables pour construire et entraîner les modèles eux-mêmes. Si cela arrive, les futures versions de Claude pourraient être continuellement améliorées par Claude lui-même.

Les preuves venues du monde extérieur

Le rythme d'amélioration des modèles d'IA s'accélère. La durée des tâches qu'ils parviennent à mener à bien de manière fiable, seuls, double environ tous les quatre mois – contre tous les sept mois auparavant. En mars 2024, Claude Opus 3 pouvait accomplir des tâches logicielles qui prennent environ quatre minutes à un humain. Un an plus tard, Claude Sonnet 3.7 gérait des tâches d'environ une heure et demie. Un an après, Claude Opus 4.6 gérait des tâches de douze heures.[1] Si la tendance se confirme, des tâches qui prennent plusieurs jours à une personne qualifiée pourraient devenir accessibles cette année. En 2027, les systèmes d'IA pourraient être capables de tâches qui prennent des semaines à une personne.

Le même schéma se retrouve sur les benchmarks de code et de recherche. Un benchmark mesure la performance des modèles dans un domaine donné ; on dit qu'il est « saturé » lorsque les modèles approchent les 100 %.[2] **SWE-bench** est un test de référence de génie logiciel en conditions réelles : on confie au modèle un véritable dépôt de code open source et un vrai rapport de bug, et on lui demande d'écrire une modification qui corrige le problème et passe les propres tests du projet. Les modèles sont passés de scores à un chiffre à la saturation du benchmark en deux ans.

CORE-Bench teste la capacité d'un modèle à reproduire des recherches existantes – un prérequis pour qu'il mène un jour des recherches originales. On lui fournit le code et les données d'un article publié, et on lui demande de tout réexécuter et de confirmer qu'il reproduit les résultats. Les systèmes d'IA sont passés d'environ 20 % de réussite en 2024 à la saturation du benchmark quinze mois plus tard.

METR, l'organisme qui fait tourner le benchmark mesurant la capacité des modèles à accomplir des tâches de longue durée, a constaté que Claude Mythos Preview pouvait travailler « au moins » 16 heures, « à la limite supérieure de ce que [METR] peut mesurer sans nouvelles tâches ».

Les benchmarks publics en disent long sur les capacités de ces systèmes. Mais ils ne révèlent pas l'effet des systèmes d'IA sur l'accélération du développement de l'IA elle-même. Pour cela, il faut des preuves directes, internes aux entreprises d'IA comme Anthropic.

Les preuves venues de l'intérieur d'Anthropic

Construire un modèle de pointe suppose deux grandes catégories de travail. Il y a l'*ingénierie* : écrire le code, monter l'infrastructure, superviser l'entraînement du modèle. Et il y a la *recherche* : décider quelles expériences mener, interpréter les résultats, déterminer quelles idées essayer ensuite.

Sur l'ingénierie comme sur la recherche, le constat est cohérent. En ingénierie, on peut confier à Claude un problème sous-spécifié et il trouve comment le résoudre ; les humains fournissent l'objectif, mais n'ont plus besoin de fournir la méthode. En recherche, Claude égale ou surpasse déjà des humains qualifiés pour exécuter une expérience bien spécifiée. En revanche, des écarts de performance importants subsistent dès qu'il s'agit, pour Claude, d'exercer un jugement dans le choix des objectifs – en ingénierie comme en recherche. C'est précisément l'écart entre l'IA d'aujourd'hui et un futur système capable de concevoir lui-même son successeur.

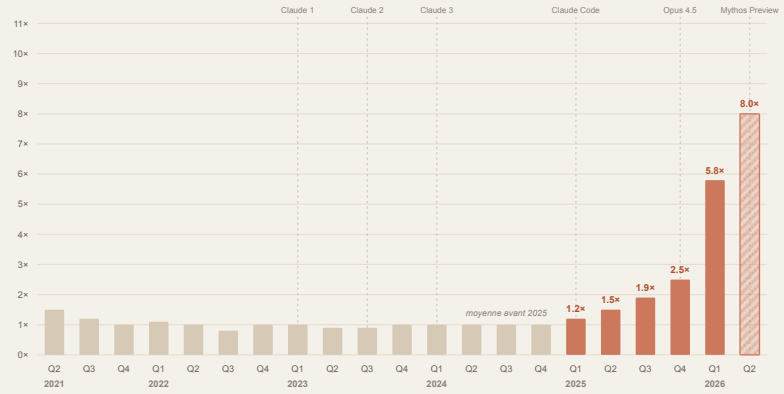
Chez Anthropic, il est courant que les employés se voient confier des tâches plus ouvertes et plus importantes à mesure qu'ils gagnent en expérience. Au début, ils exécutent une tâche

définie par quelqu'un d'autre : *« Le bouton d'export ne marche pas, peux-tu le réparer ? »* Avec l'expérience, on leur confie un objectif et ils en conçoivent eux-mêmes l'approche : *« Cherche pourquoi le réseau ralentit en cas de forte charge. »* Aux niveaux les plus seniors, ils décident des problèmes qui valent la peine d'être traités : *« Que devrait construire l'équipe au prochain trimestre ? »* Les données internes d'Anthropic permettent de voir jusqu'où Claude est allé dans sa capacité à gérer ces différents types de tâches.

Claude écrit une part importante du code d'Anthropic

En mai 2026, plus de **80 % du code** que nous intégrons à la base de code d'Anthropic était écrit par Claude.[3] Avant le lancement de Claude Code en préversion de recherche en février 2025, ce chiffre se comptait en unités de pourcent. Ce basculement se voit aussi dans la production par ingénieur. Le nombre de lignes de code intégrées par ingénieur et par jour est resté constant pendant les quatre premières années d'Anthropic (2021-2024), puis a commencé à grimper en 2025, quand Claude s'est mis à exécuter le code au lieu de seulement le suggérer pour qu'un ingénieur le copie-colle. La pente s'est encore accentuée en 2026, lorsque les modèles ont commencé à travailler de façon autonome sur des horizons de temps plus longs. Ces deux points d'inflexion apparaissent sur le graphique ci-dessous. Au deuxième trimestre 2026, l'ingénieur type intégrait **8× plus de code** par jour qu'en 2024.[4] La raison : l'essentiel du code est écrit par Claude, l'ingénieur dirigeant et relisant plutôt que tapant lui-même.

Code contribué par personne, par trimestre



Chaque barre est la moyenne, sur les jours du trimestre, des lignes de code intégrées par contributeur actif — exprimée en multiple de la moyenne d'avant 2025. La dernière barre (hachurée) est un trimestre partiel. Les lignes pointillées marquent des dates d'annonce publique.

Une mise en garde : le nombre de lignes de code est une mesure imparfaite, car elle privilégie la quantité sur la qualité. Donc *8× lignes/ingénieur/jour* au deuxième trimestre 2026 surestime presque certainement le véritable gain de productivité. Cela indique néanmoins une accélération. Chez Anthropic, nous ne récompensons pas les gens pour le nombre de lignes qu'ils écrivent ; s'ils produisent plus de code, c'est simplement parce qu'ils utilisent des systèmes d'IA pour en écrire davantage.

La hausse du volume de code rejoint l'impression subjective de fortes hausses de productivité. Lors d'un sondage de mars 2026 auprès de 130 employés des équipes de recherche d'Anthropic, le répondant médian estimait produire environ **4× plus de résultats** avec Mythos Preview que sans accès à aucun modèle d'IA, sur le type de projets qu'il aurait menés de toute façon.[5] Nous pensons que le gain réel en mars était sans doute un peu plus faible.[6] Le constat d'ensemble nous paraît néanmoins plausible et cohérent avec nos autres observations : une fraction importante du personnel technique d'Anthropic accomplit son cœur de métier plusieurs fois plus vite qu'elle ne le pourrait sans assistance de l'IA.

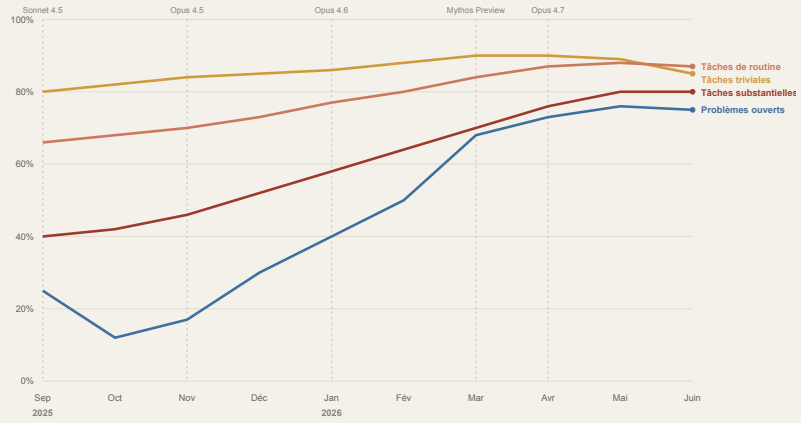
Nous constatons aussi que les gens chez Anthropic utilisent Claude pour faire un travail qui, autrement, n'aurait tout simplement pas eu lieu : construire des outils exploratoires, traiter des nettoyages longtemps repoussés. Par exemple, en avril 2026, Claude a livré plus de 800 correctifs qui ont réduit d'un facteur mille une catégorie d'erreurs d'API. L'ingénieur qui supervisait Claude a estimé qu'un humain aurait mis quatre ans à accomplir ce travail ; corriger les bugs des autres est lent et pénible, et les humains peinent à garder en tête autant de contexte qui ne leur est pas familier.

> “J'ai commencé à « claudifier » à fond il y a environ un an. Ça a été une aventure folle, et ça fait maintenant ~5 mois que je n'ai pas écrit une ligne de code moi-même.
>
> — Un employé d'Anthropic

Le code que Claude écrit est « bon » et s'améliore

« Bon code » veut dire deux choses : il fonctionne, et il est écrit de façon qu'un autre ingénieur puisse le comprendre et bâtir dessus. Sur le premier critère, la preuve est nette. Le taux auquel le personnel d'Anthropic corrige, réoriente ou reprend la main en cours de tâche face à Claude baisse régulièrement depuis un an – y compris sur les tâches les plus complexes et ouvertes. Il s'agit des problèmes sans spécification claire, où l'ingénieur ne sait pas lui-même à quoi ressemble la réponse. Cela ressort du taux de réussite de Claude au fil du temps sur des tâches de difficulté variable, comme le montre le graphique ci-dessous. Claude écrit du code qui marche.

Taux de réussite des sessions Claude Code



Sessions Claude Code internes chez Anthropic. Hebdomadaire, moyenne glissante sur 4 semaines, réussite jugée par un modèle. La complexité de la tâche est attribuée par un classifieur qui résume chaque session et la range dans l'un des quatre niveaux d'un barème fixe.

Comment lire ce graphique : *la réussite d'une session est déterminée par un Claude juge ; une session est jugée réussie si l'agent Claude Code a clairement accompli les tâches de l'utilisateur sans nécessiter de corrections. Les variations de charge de travail peuvent entraîner des fluctuations de court terme.*

Sur les tâches les plus ouvertes, le taux de réussite de Claude a atteint **76 % en mai 2026**, en hausse de 50 points en six mois. Un exemple de tâches de ce niveau : une mise à jour de routine s'est mise à faire planter des dizaines de milliers de tâches d'entraînement. Un ingénieur a pointé Claude sur l'incident en cours avec à peine quelques éléments de contexte écrit et un accès au cluster. En passant en revue les tâches en cours et en testant un paramètre d'environnement à la fois, Claude a isolé l'unique drapeau de débogage obscur qui déclenchait le plantage, l'a reproduit de façon fiable et a confirmé un correctif. En environ deux heures, Claude a livré ce qui représenterait normalement deux à trois jours de travail.

Le second critère, c'est d'écrire du code qu'un autre ingénieur peut comprendre et sur lequel il peut bâtir. Là, l'écart entre humains et IA persiste, mais se resserre vite. Il n'y a pas de consensus complet parmi le personnel d'Anthropic, mais beaucoup estiment que le code écrit par Claude était encore de moins bonne qualité que le code humain chez Anthropic fin 2025, et qu'il est aujourd'hui à parité. Nous nous attendons à ce qu'il soit meilleur d'ici un an.

Cela a changé la façon dont Anthropic relit désormais son propre code. Les modifications proposées à notre base de code sont maintenant lues par un relecteur Claude automatisé qui traque bugs, failles de sécurité et autres défauts avant toute intégration. Grâce à cet outil, nous avons mené une analyse rétrospective : une relecture Claude automatisée de chaque modification de notre base de code aurait détecté environ un tiers des bugs à l'origine d'incidents passés sur claude.ai avant même qu'ils n'atteignent la production. Les ingénieurs qui ont écrit ce code comptent parmi les meilleurs au monde pour bâtir ces systèmes. Claude rattrape aujourd'hui les erreurs qui leur avaient échappé.

> “Le code écrit par Claude était un peu moins bon que le code humain chez Anthropic fin 2025, il est à parité aujourd'hui, et nous nous attendons à ce qu'il soit strictement meilleur d'ici un an.
>
> — Anthropic

Claude est bon pour mener des expériences vers un objectif fixé par quelqu'un d'autre

À chaque sortie de modèle, Anthropic fait passer le même test : nous donnons à Claude du code qui entraîne un petit modèle d'IA, et lui demandons de le faire tourner le plus vite possible tout en passant les mêmes contrôles de validité. L'objectif et les critères de réussite sont fixés d'avance ; le travail de Claude consiste donc à trouver des accélérations en réécrivant le code, en l'exécutant, en le chronométrant, et en recommençant. C'est une version miniature d'une boucle de recherche expérimentale. En mai 2025, Claude Opus 4 obtenait en moyenne une accélération d'environ **3×** par rapport au code de départ. En avril 2026, Claude Mythos

Preview atteignait **~52×**. Pour situer, un chercheur humain qualifié aurait besoin de quatre à huit heures pour atteindre 4×.[7] Sur cette partie du travail de recherche — optimiser des étapes au sein d'une expérience clairement définie — Claude est passé de très utile à surhumain en moins d'un an.

> “La forme que prend le travail aujourd'hui, c'est en gros : « les humains ont les idées, et les modèles sont capables de les implémenter, de les tester et de les évaluer un ordre de grandeur plus vite qu'avant. »
>
> — Un employé d'Anthropic

Claude est de plus en plus capable de proposer ses propres expériences

En avril 2026, Anthropic a publié la première démonstration de Claude menant un projet de recherche ouvert de bout en bout. Des agents pilotés par Claude se sont vu confier un problème ouvert de sûreté de l'IA — en gros, *un modèle plus faible peut-il superviser de façon fiable un modèle plus fort ?* — et ont été laissés à eux-mêmes pour le résoudre. Cela impliquait de formuler des hypothèses, de les tester, de partager les résultats avec des agents en parallèle, et d'itérer. La tâche a un « plancher » et un « plafond » de performance clairs : le plancher, c'est ce que ferait le superviseur faible seul ; le plafond, c'est la performance du modèle fort entraîné sur les bonnes réponses. Deux chercheurs humains, en environ une semaine, ont récupéré à peu près **23 %** de cet écart ; les agents en ont récupéré **97 %** sur 800 heures cumulées, pour environ 18 000 $ de calcul. Ce travail comporte des réserves : le résultat ne s'est pas transféré proprement aux modèles à l'échelle de la production, et ce sont des humains qui ont choisi le problème et créé la grille d'évaluation. Mais dans ces

limites, les agents ont conçu eux-mêmes chacune des expériences. Fixer la direction a été le seul rôle vraiment joué par un humain.

> “Claude a fait tout cela avec une aide assez minimale de ma part, en l'espace de 1 à 2 jours. Si [un collègue junior] revenait vers moi avec de tels résultats dans le même laps de temps, je serais plutôt impressionné. Le futur, c'est maintenant.
>
> — Un employé d'Anthropic

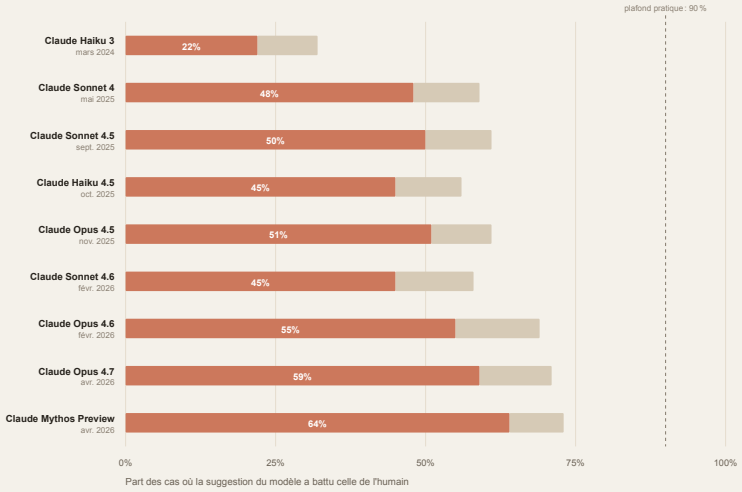
Claude est de plus en plus capable d'orienter les sessions de recherche vers des résultats

Nous avons examiné de vraies sessions Claude Code (entre janvier et mars 2026) où des chercheurs d'Anthropic travaillaient avec Claude sur un problème d'investigation ouvert : comprendre pourquoi un entraînement plantait sans cesse, ou pourquoi un modèle obtenait un mauvais score sur un benchmark. Dans chaque cas, nous avons repéré un moment où le chercheur a fait un détour : il a suivi une piste qui a fait dériver la session avant qu'elle ne revienne sur les rails. Nous avons alors montré à différents modèles Claude *uniquement* le travail antérieur au moment où la session a déraillé, en leur demandant ce qu'ils feraient ensuite. Un autre Claude, capable de voir comment la session a fini par tourner, jugeait ensuite si la proposition de l'IA ou celle de l'humain était la meilleure.[8]

Comme nous avons délibérément choisi des moments (n=129) où le choix de l'humain était perfectible, ce n'est pas une comparaison à armes égales entre le jugement du modèle et celui de l'humain. Ce que ces moments offrent, c'est un

ensemble de situations réalistes et difficiles, où le bon prochain pas n'est pas évident, et où le choix de l'humain sert de repère utile pour comparer la performance des modèles dans le temps. Sur cette mesure, notre meilleur modèle de novembre 2025 (Opus 4.5) battait le choix humain **51%** du temps ; en avril 2026 (Mythos Preview), ce chiffre est monté à **64%**. Le travail quotidien de recherche est en grande partie une chaîne de ces décisions de « prochain pas », ce qui en fait une mesure pertinente de la capacité du modèle à mener un jour une investigation de lui-même. Nous voyons ce résultat comme un signal précoce que les systèmes d'IA progressent dans le type de jugement dont dépend la recherche en IA.

Là où un chercheur s'est trompé, Claude aurait-il fait mieux ?



L'étude porte sur 129 sessions de recherche Claude Code internes, chacune coupée à un moment où la direction choisie par l'humain était perfectible. Le modèle propose une alternative ; un juge qui connaît les conclusions finales de la session désigne le meilleur choix. La barre claire indique la part d'égalités.

Comment lire ce graphique : *la ligne de « plafond pratique » mesure une réponse « idéale » rédigée par un modèle ayant pu voir l'ensemble de la session (y compris sa conclusion).*

> “L'avantage comparatif des humains, à l'heure actuelle, reste de voir le tableau d'ensemble et de penser au-delà des limites de la tâche immédiate.
>
> — Un employé d'Anthropic

À quoi pourrait ressembler l'avenir du travail chez Anthropic ?

Les preuves suggèrent que le rôle humain se rétrécit à chaque étape du processus de développement de l'IA. Une fois que la qualité du code écrit par l'humain et par l'IA atteint la parité, les humains cesseront tout à fait d'écrire du code et se contenteront de le relire. Mais s'ils ne peuvent pas le relire aussi vite que Claude le génère, la relecture humaine deviendra le goulot d'étranglement du développement de l'IA. De même, dès lors que Claude peut mener des expériences, la question se déplace vers : « Lesquelles valent la peine d'être menées ? » En clair : *faire* (écrire le code, mener l'expérience, produire le résultat) ne coûte presque plus de temps humain, même si cela a encore un coût en calcul.

Un domaine d'avantage comparatif humain, pour l'instant, c'est le goût et le jugement en recherche : choisir quels problèmes comptent, à quels résultats se fier, et quand une approche est une impasse.

> “Le travail (et la vie) reposait sur une économie du don, faite de petits services entre humains. « Tu peux m'aider à faire tourner ce script ? » [...] chacun créait une petite dette, une petite conscience mutuelle. [Claude est] plus rapide, il ne crée aucune dette — mais chacun de ces échanges est une occasion perdue de collaboration humaine.
>
> — Un employé d'Anthropic

> “Les jours où tout fonctionne bien, je ne peux m'empêcher de penser que rien de ce que je fais n'a d'importance : tout est automatisé, en mieux et plus vite que je ne le serai jamais. Mais il y a aussi les jours où tout casse, sans que je comprenne pourquoi, et je réalise que je n'ai plus la moindre idée de ce que j'ai fabriqué.
>
> — Un employé d'Anthropic

Et si nous avions tort ?

Une objection naturelle aux preuves présentées plus haut : ce qui reste entre les mains des humains — choisir les problèmes à traiter — serait justement ce qui compte le plus. Sans ce jugement, Claude est un assistant compétent, mais pas un système capable de faire avancer l'IA par lui-même.

On ne sait pas vraiment si les méthodes d'entraînement et les architectures actuelles peuvent débloquer cette capacité. Mais l'IA progresse rarement par moments « eurêka ». Il y en a eu quelques-uns dans l'histoire récente de l'IA — l'architecture Transformer, ou les modèles à mélange d'experts — mais les idées qui changent de paradigme arrivent à des années d'intervalle. Entre les deux, l'essentiel du progrès est incrémental : on passe quelque chose à l'échelle, on regarde ce qui casse, on corrige, et on recommence. C'est exactement le type de travail où Claude excelle désormais. Edison disait que le génie, c'est 1 % d'inspiration et 99 % de transpiration. Or nous voyons la transpiration s'automatiser de plus en plus. Il devient clair qu'une grande part de ce qui fait avancer la frontière est automatisable ; le progrès de la recherche à grande échelle est surtout fonction des outils et des ressources, qui déterminent la vitesse à laquelle on peut mener des expériences, le nombre qu'on peut lancer en parallèle, et la rapidité avec laquelle on obtient des résultats.

Même en supposant que Claude n'acquière jamais un bon goût de recherche, une lecture prudente de nos données implique tout de même une accélération cumulative. Si les humains passent l'essentiel de leur temps sur la fraction infime du travail qu'est la définition de la direction, pendant que Claude se charge du reste, alors chaque ingénieur ou chercheur pilote bien plus de travail qu'auparavant. Les données que nous observons suggèrent que les gens chez Anthropic vont à la fois plus vite et couvrent un terrain plus large. En pratique, cela signifie que l'IA fait déjà avancer Anthropic bien plus vite qu'avant l'arrivée d'outils d'IA efficaces.

La lecture moins prudente, c'est que les premières données sur l'amélioration du jugement de recherche de Claude — aussi limitées soient-elles aujourd'hui — indiquent que cette

capacité progresse elle aussi. Le « goût de recherche » pourrait n'être qu'une capacité de plus que les systèmes d'IA échouent à maîtriser un temps, avant d'y devenir bons. Nous avons observé le même schéma pour d'autres compétences qualitatives : expliquer pourquoi une blague est drôle, faire preuve de théorie de l'esprit, résoudre des énigmes linguistiques.

Futurs possibles

Ce qui se passera ensuite dépend de deux choses : la tendance va-t-elle se poursuivre, et que choisirons-nous de faire si c'est le cas. On peut imaginer au moins trois scénarios.

1 **La tendance plafonne, mais les capacités actuelles de l'IA se diffusent largement.** Cet article regorge de trajectoires exponentielles. Mais ces trajectoires pourraient en réalité être des courbes en S. Nous approchons peut-être du coude de la courbe, là où les rendements d'échelle diminuent, où la ligne se redresse puis s'aplatit. Le jugement qui sépare un chercheur compétent d'un grand chercheur pourrait être une capacité qui ne naît pas de la mise à l'échelle d'intrants comme le calcul et les données. Si c'est le cas, franchir ce goulot exigerait une idée nouvelle, comme une approche architecturale qui supplanterait le Transformer sur lequel reposent tous les modèles de pointe actuels.

Autre possibilité : la contrainte qui borne le progrès de l'IA pourrait se situer dans la chaîne d'approvisionnement, pas dans le modèle — faire avancer et diffuser la frontière pourrait exiger plus d'énergie et de calcul qu'il n'en existe aujourd'hui. Le rythme de fabrication des puces, l'expansion du réseau électrique ou la bande passante d'interconnexion pourraient être la contrainte, plutôt que l'intelligence elle-même. On ne peut pas non plus exclure un choc exogène qui ralentirait brutalement les choses — une

chute soudaine de l'offre de calcul ou d'électricité, qui freinerait le progrès et rendrait l'investissement des labos plus coûteux. Ou bien un autre obstacle que nous n'anticipons pas.

Même si les capacités des modèles étaient gelées au niveau actuel, nous nous attendrions à des changements majeurs dans le monde. Le projet Glasswing en est un signe précoce : dès ses premières semaines, Mythos Preview a trouvé plus de dix mille vulnérabilités logicielles de gravité haute ou critique dans les systèmes les plus importants du monde — au point que le goulot d'étranglement de la cyberdéfense est déjà passé de la détection des vulnérabilités à leur correction assez rapide. Et nous n'en sommes qu'au début de la diffusion des modèles actuels dans l'économie, où une entreprise de 100 personnes pourra de plus en plus faire le travail d'une entreprise de 1 000, parce que chaque employé sera assis au sommet d'une pyramide d'agents.

Nous incluons ce scénario par souci d'exhaustivité, mais nous ne le jugeons pas probable. Toutes les capacités que nous savons mesurer — y compris les plus « floues », comme la qualité du code et la réussite sur les tâches ouvertes — ont jusqu'ici suivi la même courbe. Nous n'avons pas encore vu cette courbe s'infléchir. Des trois futurs envisagés, celui-ci laisserait aux gouvernements et aux sociétés le plus de temps pour s'adapter. Ce sont les deux suivants qui nous inquiètent davantage : ils iraient plus vite et laisseraient bien moins de marge pour se préparer.

2 **Les labos d'IA continuent d'engranger des gains d'efficacité cumulatifs.** Dans ce scénario, le développement de l'IA s'automatise largement, mais les humains continuent de fixer les directions de recherche et de juger les résultats. Les organisations qui utilisent des systèmes d'IA deviendraient bien plus

efficaces avec le temps, au point qu'on pourrait voir des multiplicateurs de productivité importants sur chaque personne. Des entreprises de 100 personnes pourraient faire le travail d'organisations de 10 000 ou 100 000. Cela révolutionnerait le travail intellectuel et les services publics, mais pourrait aussi être détourné à des fins néfastes — de la surveillance autoritaire de populations entières aux opérations d'influence qui adaptent la manipulation à chaque individu, à une échelle qu'aucune équipe humaine ne pourrait égaler. Le rôle des humains dans des entreprises comme Anthropic se déplacerait : les gens s'associeraient à des systèmes d'IA pour démultiplier la recherche et générer de nouvelles idées, et construiraient ensemble les systèmes nécessaires pour vérifier que les sorties de l'IA sont dignes de confiance.

Les preuves que nous avons exposées suggèrent que nous nous dirigeons probablement vers ce scénario. Mais accélérer une partie d'un processus ne fait souvent que déplacer le goulot d'étranglement ailleurs : le rythme global est plafonné par les parties qui n'ont pas accéléré. En informatique, c'est la *loi d'Amdahl*, et la même logique vaut pour les organisations. Anthropic a déjà rencontré une signature de la loi d'Amdahl : à mesure que nous faisons circuler plus de code, la relecture humaine est devenue un nouveau goulot.

Nous avons aussi rencontré cette friction hors de l'ingénierie. Il y a eu une explosion de nouvelles idées, initiatives, outils et simulations, fruit du travail des employés d'Anthropic avec des modèles très capables — bien plus que ce que nous avons la capacité de poursuivre. La vitesse à laquelle une organisation sait repérer et lever ces goulots est peut-être une compétence qui s'améliore avec le temps, et pourrait devenir la compétence la plus importante pour toute organisation.

3 **Les systèmes d'IA deviennent eux-mêmes capables d'une auto-amélioration récursive complète et commencent à construire leurs successeurs.** Si les tendances techniques d'amélioration des capacités se poursuivent, *et* que les systèmes d'IA parviennent à développer les capacités propres à l'ingéniosité humaine transformatrice, alors il est plausible qu'ils puissent se concevoir et se perfectionner eux-mêmes.

Dans ce monde, le rythme du progrès serait entièrement déterminé par la disponibilité du calcul (ou par la vitesse à laquelle on découvre diverses efficacités d'entraînement ou d'inférence). Les humains joueraient un rôle nettement réduit, déplaçant l'essentiel de leurs efforts vers la supervision, la validation et la vérification d'un « laboratoire virtuel » en expansion, piloté par des systèmes d'IA. Nous nous attendons à ce que des systèmes capables de recherche et développement d'IA automatisés possèdent des compétences transférables au reste de la science, leur permettant de commencer à révolutionner d'autres domaines.

La manière dont le problème de l'alignement se résoudra – ou non – dans ce futur est ce sur quoi nous sommes le moins certains. Les modèles pourraient se révéler suffisamment alignés et dotés d'assez de goût de recherche pour découvrir et mettre en œuvre des solutions inédites. Ils pourraient aussi être assez sages pour interrompre le développement si nécessaire. À l'inverse, les rares cas de désalignement présents dans les modèles actuels pourraient se cumuler à mesure que les modèles construisent leurs successeurs, devenant plus fréquents mais moins compris, jusqu'à ce que nous en perdions le contrôle. Il est possible que nous ne parvenions pas à construire, intégrer et vérifier les outils dont nous aurions besoin pour comprendre sur quelle trajectoire nous nous trouvons réellement.

Nous n'avons pas de bonne intuition de ce à quoi ce monde ressemblerait, car notre économie est aujourd'hui mue par les humains et les outils qu'ils ont bâtis. Par nature, un monde mû par une auto-amélioration récursive rapide pourrait finir dominé par le modèle qui s'auto-améliore, à mesure que ses capacités éclipsent celles des humains et qu'il prolifère dans l'économie. Il est difficile de prédire à quoi ressemble l'économie si le travail humain cesse d'être compétitif.

Même si le développement des modèles devenait entièrement automatisé et récursif, nous ne pouvons pas prédire ce que cela signifierait pour la vie quotidienne de la plupart des gens. La loi d'Amdahl s'applique ici aussi. Une intelligence récursive pourrait permettre d'atteindre rapidement, dans certains domaines, nombre des bienfaits décrits dans *Machines of Loving Grace*. Nous nous attendons à ce que l'intelligence incarnée (la robotique) suive de près l'intelligence récursive. Une intelligence plus puissante pourrait nous aider à construire plus vite dans le monde physique, à mener des essais cliniques plus productifs de médicaments qui sauvent des vies, et à développer de nouvelles formes de coordination.

Mais atteindre l'auto-amélioration récursive ne suppose pas, à soi seul, un changement immédiat dans la façon dont la production industrielle s'organise, dont les sociétés se structurent, ou dont les marchés fonctionnent. Plus d'intelligence ne peut pas apprendre ce que fait un médicament au fil de décennies d'usage, ni tenir des élections plus tôt qu'une constitution ne le prévoit, ni transformer un inconnu en vieil ami le temps d'un week-end. Pour la plupart des gens, le rythme ressenti de ce futur restera fixé par les goulots d'étranglement, même si le laboratoire en amont tourne à la vitesse du calcul.

> Cette collision – une intelligence récursive se construisant elle-même toujours plus vite, qui rencontre le monde des humains, des relations et de la gouvernance – est une autre part de ce futur que nous ne pouvons pas prédire.

Que devrions-nous faire ?

S'il était possible de ralentir efficacement le développement de cette technologie pour nous donner plus de temps face à ses immenses implications, nous pensons que ce serait probablement une bonne chose. Mais si un ralentissement ne fait que permettre aux acteurs les moins prudents de rattraper leur retard, il pourrait rendre tout le monde moins sûr. Sans mécanisme de coordination mondiale, entreprises et gouvernements devront prendre des décisions difficiles sur la sécurité, sous la pression de la concurrence et de la géopolitique.

Nous croyons qu'il serait bon pour le monde d'avoir l'*option* de ralentir ou de suspendre temporairement le développement de l'IA de pointe, afin que les structures sociétales et la recherche sur l'alignement puissent suivre le rythme de la technologie. L'Anthropic Institute mènera des recherches – en collaboration avec beaucoup d'autres – et agira pour aider à construire les systèmes qu'un ralentissement ou une pause crédibles exigeraient. Ces systèmes permettraient aux développeurs d'IA de pointe de vérifier que les autres, partout dans le monde, ont effectivement arrêté ou ralenti, et qu'un acteur malveillant ne pourrait pas profiter d'un ralentissement coordonné pour prendre secrètement de l'avance. Si de tels systèmes existaient, nous nous attendons à ralentir ou suspendre temporairement, à condition que les autres développeurs à la frontière ou proches d'elle en fassent autant, de façon vérifiable.

Un ralentissement ou une pause significatifs exigeraient que plusieurs labos bien dotés, à la frontière ou proches d'elle, dans plusieurs pays, acceptent de s'arrêter aux mêmes

conditions. Il faudrait aussi que chacun puisse vérifier que les autres se sont effectivement arrêtés. En raison des caractéristiques propres aux systèmes d'IA, l'élément de détectabilité (un critère moins exigeant que la vérifiabilité) de ce problème de contrôle des armements est bien plus délicat qu'avec d'autres technologies. Les entraînements sont bien plus faciles à dissimuler que des silos à missiles, leurs intrants sont à usage général, et l'incitation à tricher discrètement est énorme, car celui qui poursuit pendant que les autres s'arrêtent pourrait hériter de la tête de course. Une pause crédible doit aussi préciser ce qui la déclenche, ce qui la lève, et qui en est l'arbitre.

Rien de tout cela n'est nécessairement impossible en principe – le monde a su bâtir des régimes de vérification pour d'autres technologies complexes (par exemple le traité sur les forces nucléaires à portée intermédiaire) – mais ces régimes ont mis des décennies à construire l'infrastructure et la confiance. Nous n'avons pas autant de temps. Une pause unilatérale par un seul labo, à l'inverse, est réalisable immédiatement, mais accomplit bien moins : elle changerait l'identité du meneur, sans créer le processus délibératif plus large qui manque aujourd'hui.

Dans les mois à venir, nous organiserons des échanges où responsables politiques, chercheurs, société civile et autres entreprises d'IA pourront aider à répondre à certaines des questions soulevées ici, en particulier autour de l'auto-amélioration récursive complète et de la façon de créer de meilleures options de coordination et de délibération. Nous publierons ce qui en ressortira. La fenêtre pour explorer ces questions ensemble est là, et des personnes extérieures aux entreprises d'IA devraient prendre part à cette délibération.

Quand l'IA se construit elle-même — traduction française de l'essai « When AI builds itself » de l'Anthropic Institute. Les données, graphiques et citations reproduisent fidèlement l'original ; les graphiques ont été redessinés pour cette édition française. Les appels de note (1–8) sont ceux de l'original.

Traduction et mise en page : **Diederick Legrain** · AI Shift · ai-shift.be