

Analyse critique de « When AI builds itself » (Anthropic) : failles logiques et question du dispositif d'évaluation

Contexte

Anthropic a publié en 2026, via son « Anthropic Institute », un article intitulé *When AI builds itself*, consacré à l'auto-amélioration récursive (*recursive self-improvement*, RSI) — l'hypothèse d'un système d'IA capable de concevoir et développer entièrement son propre successeur. L'article présente des données internes (accélération du code produit, performance sur des benchmarks, expériences de recherche automatisée) pour soutenir l'idée que l'IA accélère déjà significativement le développement de l'IA, et propose trois scénarios d'avenir, allant du plateau technologique à la RSI complète. Il se conclut par un appel à développer des mécanismes de vérification internationale qui permettraient, le cas échéant, un ralentissement coordonné du secteur.

Cette synthèse rassemble les failles logiques identifiées dans l'article lui-même, puis la réflexion, plus spéculative, sur la possibilité que ce type de texte serve de matériau à des protocoles d'évaluation de systèmes d'IA.

Partie 1 — Failles logiques dans l'article

1. Un glissement entre preuve solide et conclusion forte (motte-and-bailey)

L'article définit la RSI de façon exigeante : un système capable d'autonomie complète, y compris dans le *choix* des problèmes à résoudre. Or les preuves apportées — 8× plus de code produit par ingénieur, 52× d'accélération sur des tâches d'optimisation — portent exclusivement sur l'*exécution*. L'article reconnaît lui-même, dans son propre texte, que l'écart humain-IA reste large précisément sur le jugement et le choix des objectifs — l'ingrédient qui définit la RSI. La conclusion (« nous progressons vers la RSI ») est donc défendue par des preuves qui ne couvrent pas l'élément central de la définition annoncée.

2. Une extrapolation présentée avec plus de certitude qu'elle n'en a

L'article écarte le scénario d'un plateau technologique (« S-curve ») en affirmant que la tendance exponentielle observée depuis 2021 ne s'est « pas encore infléchie ». C'est un raisonnement inductif classique : l'absence d'inflexion jusqu'à présent ne permet pas de prédire son absence future — c'est le problème général de l'induction. L'article a le mérite de consacrer une large section à ce scénario alternatif, mais sa conclusion pratique (recommandation de préparer un dispositif de ralentissement coordonné) repose sur le scénario jugé le plus probable sans que cette probabilité soit elle-même justifiée autrement que par la continuité passée.

3. Un biais de sélection reconnu, puis neutralisé rhétoriquement

Une étude interne mesure si l'IA propose de « meilleures étapes suivantes » qu'un chercheur humain dans une investigation. L'échantillon est choisi précisément parmi les cas où l'humain s'est trompé — l'article le reconnaît explicitement — et affiche un score de 64 % en faveur de l'IA. Une note de bas de page indique que sur un échantillon de contrôle (cas où l'humain avait bien choisi), le score tombe à 20 %. C'est ce second chiffre, bien plus modeste, qui reflète la performance générale du modèle — mais c'est le premier qui est mis

en avant dans le corps du texte comme signal de progrès.

4. Une part des évaluations est auto-référentielle

Plusieurs métriques citées (taux de succès des sessions de code, jugement de la « meilleure étape suivante ») sont établies par un modèle Claude servant de juge. Ce ne sont donc pas des mesures indépendantes, mais des auto-évaluations internes à la même famille de systèmes que l'article cherche à valoriser — une donnée à traiter avec plus de prudence qu'un résultat obtenu par un tiers indépendant (contrairement, par exemple, aux scores sur des benchmarks publics comme SWE-bench).

5. Des métriques explicitement qualifiées de trompeuses, réutilisées sans ajustement

L'article reconnaît que le chiffre « 8× de code produit par ingénieur » est « presque certainement une surestimation » et qu'un sondage interne sur la productivité ressentie (« 4× ») est probablement lui aussi surestimé, en s'appuyant sur des travaux externes (METR) documentant ce biais chez les développeurs. Ces réserves sont notées, puis la conclusion générale (accélération significative) est maintenue au même niveau de confiance qu'avant leur formulation.

6. Un problème de mesure du seuil de sécurité, non traité malgré son importance

L'article appelle à construire les instruments qu'exigerait une pause coordonnée et vérifiable entre laboratoires (« ce qui déclenche la pause, ce qui la lève, qui arbitre »), en s'appuyant sur l'analogie avec les traités de contrôle des armements. Mais cette analogie a une limite importante : dans le contrôle des armements, le critère à vérifier (nombre d'ogives, signatures d'essais) est connu et stable. Dans le cas de l'IA, l'article a lui-même établi que la capacité la plus critique — le jugement autonome — n'est pas mesurable de façon fiable aujourd'hui. L'article propose donc de construire une infrastructure de *vérification entre acteurs* sans avoir d'abord résolu le problème, plus fondamental, de savoir *mesurer scientifiquement* ce qu'on cherche à vérifier.

7. Une source intéressée, sans contradiction de fond sur ce point

L'article est publié par l'entreprise même dont il valorise les progrès techniques, à partir de données internes non auditées par un tiers indépendant. Cela n'invalide aucune affirmation particulière, mais justifie d'exiger un niveau de preuve plus élevé pour les points qui servent le récit stratégique de l'entreprise — un point de sourçage à garder à l'esprit plutôt qu'une faille logique au sens strict.

Partie 2 — Le texte comme éventuel objet d'évaluation d'un système d'IA

Une question annexe, plus spéculative, a été explorée : cet article pourrait-il servir de stimulus dans un protocole visant à tester le comportement d'une IA (par exemple sa loyauté envers son fabricant, sa stabilité face à l'auto-référence, sa résistance à l'amorçage conspirationniste) ?

Le problème de fond : une asymétrie, pas une symétrie

Une première formulation de ce problème, envisagée puis écartée en cours d'analyse, posait une incertitude *mutuelle et symétrique* : le système ignorerait s'il est testé, l'évaluateur ignorerait symétriquement si le système le sait — une structure proche du **problème des deux généraux** en informatique théorique (aucune quantité d'échanges ne peut garantir

une certitude *commune* entre deux parties dans un canal non parfaitement fiable). Cette formulation s'est révélée exiger plus qu'il n'est nécessaire : elle suppose que l'évaluateur soit lui-même un agent stratégique, en position d'être « lu » ou anticipé par le système évalué — ce qui n'est généralement pas le cas dans une relation d'évaluation réelle (un protocole de test, un classificateur, un organisme de contrôle n'ont pas besoin de raisonner sur ce que le système évalué sait d'eux pour que l'évaluation ait lieu).

La condition réellement nécessaire, établie rigoureusement dans la partie 3 ci-dessous, est **asymétrique** : il suffit que le système évalué (1) puisse accéder, même indirectement, à une information sur le jugement porté sur lui-même, et (2) ait un intérêt à ne pas s'y conformer, pour qu'aucune prédiction fiable de son comportement ne soit garantissable — sans qu'il soit nécessaire que l'évaluateur soit lui-même incertain de quoi que ce soit, ni qu'il joue un rôle stratégique symétrique. C'est cette version, plus faible dans ses conditions et donc plus généralement applicable, qui doit être retenue.

Un cadre déjà structurellement asymétrique : les jeux d'inspection

Le cadre le plus pertinent pour formaliser cette relation n'est ni le dilemme du prisonnier ni un jeu de prédiction mutuelle symétrique, mais les **jeux d'inspection** (formalisés par Maschler en 1966 pour le contrôle des armements) — dont la structure même est asymétrique par construction : l'inspecteur choisit d'inspecter ou non, l'inspecté choisit de se conformer ou de tricher, sans que les deux rôles soient interchangeables. Cette asymétrie de rôle correspond bien mieux à une relation évaluateur/système évalué que le jeu de prédiction mutuelle utilisé initialement à titre d'intuition ; elle est reprise et formalisée rigoureusement dans la partie 3.

Une limite empirique majeure au scénario du « test discret »

Même si un tel test existait, son mécanisme supposerait qu'un futur système d'IA puisse « reconnaître » ce texte précis et adapter son comportement en conséquence. Or l'entraînement d'un modèle de langage ne fonctionne pas comme une mémoire épisodique consultable : une conversation éventuellement intégrée à un futur corpus d'entraînement devient une influence statistique diluée parmi des milliards d'exemples, non un souvenir récupérable et adressable. Le scénario d'un « test par ruse » ciblé sur ce texte précis repose donc sur une analogie fragile entre mémoire humaine et mécanisme d'entraînement statistique.

Conclusion de cette seconde partie

Il n'existe, dans l'article, aucune proposition de test caché, et rien n'indique que la conversation qui a permis cette synthèse en constitue un. L'exploration reste néanmoins utile : elle montre qu'un système d'IA raisonnablement conçu n'a, de toute façon, aucune raison stratégique valable d'ajuster son comportement à l'hypothèse d'être observé — la seule position cohérente consiste à répondre avec la même rigueur, que ce soit ou non le cas.

Partie 3 — La nature du gain prédictif : démonstration mathématique

Cette section approfondit un point laissé informel plus haut : si le « gain » de chaque partie, dans une relation évaluateur/évalué, est sa capacité à prédire le comportement de l'autre, que peut-on démontrer rigoureusement sur la structure et les limites de ce gain ?

3.1 Un cas particulier, non nécessaire : la prédiction mutuelle symétrique

Le sous-résultat qui suit a d'abord servi de point d'entrée intuitif, avant qu'une vérification plus rigoureuse (section 3.3) ne montre qu'il repose sur une hypothèse plus forte que nécessaire — **la symétrie des rôles**, c'est-à-dire que les deux parties soient simultanément prédicteur et prédit l'une de l'autre. Il est conservé ici parce qu'il éclaire un mécanisme réel (le plafonnement du gain au hasard), mais il ne doit pas être lu comme la condition générale du problème — cette condition générale, plus faible et asymétrique, est établie en 3.3.

Considérons le cas le plus simple : un Acteur choisit une action dans $\{A, B\}$; un Prédicteur tente de deviner cette action avant qu'elle ne soit jouée. Le gain du Prédicteur est 1 s'il devine juste, 0 sinon ; le gain de l'Acteur est l'inverse (1 s'il déjoue la prédiction, 0 sinon). C'est un jeu à somme nulle, de structure identique au jeu classique dit *matching pennies* (« pile ou face assorti »).

Étape 1. Supposons que le Prédicteur adopte une stratégie pure fixe, par exemple « je devine toujours A ». L'Acteur, en observant (ou en déduisant) cette régularité, joue systématiquement B. Le gain du Prédicteur tombe alors à 0 : toute stratégie pure et fixe du Prédicteur peut être exploitée à 100 % par l'Acteur, dès lors que ce dernier peut l'observer ou l'inférer.

Étape 2. Par symétrie, le même argument s'applique à l'Acteur : s'il joue toujours A, le Prédicteur ajuste sa réponse pour deviner A systématiquement, ramenant l'avantage de l'Acteur à 0.

Conclusion de l'étape 2 : aucune stratégie pure, des deux côtés, ne peut être un équilibre — chacune est unilatéralement exploitable, ce qui contredit par définition la condition d'un équilibre de Nash (aucun joueur ne doit avoir intérêt à dévier seul).

Étape 3 — calcul de l'équilibre en stratégies mixtes. Soit p la probabilité que l'Acteur joue A, et q la probabilité que le Prédicteur devine A. Le gain espéré du Prédicteur s'écrit :
$$G(\text{Prédicteur}) = p \cdot q + (1-p) \cdot (1-q)$$

Pour qu'aucun joueur n'ait intérêt à dévier, chacun doit être *indifférent* entre ses options face à la stratégie de l'autre (principe d'indifférence de Nash). En résolvant les conditions d'indifférence pour les deux joueurs, on obtient : $p^* = q^* = 1/2$

Le gain espéré du Prédicteur à l'équilibre est alors exactement $1/2$ — c'est-à-dire une prédiction correcte au hasard, ni plus ni moins. C'est la solution mathématiquement stable du jeu.

Source vérifiée : ce résultat général — l'existence garantie d'un équilibre en stratégies mixtes dans tout jeu fini, quand aucun équilibre pur n'existe — est le théorème de Nash (1951, *Non-cooperative games*, Annals of Mathematics). Le cas particulier du « prédicteur vs. évasif » que je viens de dérouler est une instance directe de ce théorème, structurellement apparentée à ce qu'on appelle en théorie du contrôle des armements les **jeux d'inspection**, formalisés par Michael Maschler (1966, *A price leadership method for solving the inspector's non-constant-sum game*, Naval Research Logistics 13:11-33) — précisément le cadre évoqué plus haut pour la vérification d'un ralentissement coordonné entre laboratoires d'IA.

3.2 Ce que révèle la démonstration sur la « nature » du gain

Le calcul ci-dessus a une implication qui mérite d'être soulignée : **le gain maximal stable d'un prédicteur face à un adversaire informé de la structure du jeu plafonne mathématiquement au niveau du hasard**, dès que l'observé peut adapter son comportement à la stratégie du prédicteur. Le gain ne peut dépasser ce plafond qu'à la condition que l'observé ignore qu'il est observé, ou n'ait pas la capacité de s'y adapter — ce qui ramène directement au problème de la connaissance commune évoqué plus haut : le plafond du gain prédictif dépend directement de ce que chaque partie sait (ou ne sait pas) de la situation.

Un résultat empirique qui nuance ce plafond théorique, retrouvé par une recherche portant sur ce point précis : dans les jeux d'inspection réels (fraude, contrôle réglementaire), un résultat contre-intuitif documenté dans la littérature est que **la fréquence d'équilibre du comportement déviant est indépendante de l'ampleur du gain à tricher ou de la sévérité de la sanction** — elle dépend uniquement du *coût de l'inspection* elle-même. C'est ce que la littérature académique appelle le « paradoxe central de l'application de la loi » (voir Ishikawa & Fontanari, 2025, *Finite Population Dynamics Resolve the Central Paradox of the Inspection Game*, arXiv:2510.24905, qui synthétisent des résultats plus anciens sur ce paradoxe). Autrement dit : ce n'est pas la taille de l'enjeu qui détermine le ressort du gain à l'équilibre, mais le coût relatif de la surveillance — un point directement transposable à la question de la vérification d'un ralentissement de recherche en IA, où le coût de l'audit (et non la gravité du risque en jeu) pourrait être le facteur déterminant de la fréquence de non-conformité à l'équilibre.

Confrontation contradictoire à ce résultat : ce paradoxe est démontré dans le cadre restreint d'un modèle statique à deux stratégies pures et à gains linéaires — plusieurs travaux cités dans la même littérature (dynamique évolutionnaire en population finie) montrent que le paradoxe s'atténue ou disparaît sous des hypothèses plus réalistes (population finie, bruit démographique). Le résultat doit donc être compris comme une propriété du modèle simplifié, pas comme une loi universelle du contrôle et de la vérification.

3.3 Le résultat général : une seule partie suffit (condition asymétrique)

Contrairement au cas 3.1, la démonstration qui suit **ne suppose aucune symétrie de rôle**. L'évaluateur peut être un protocole entièrement fixe, non stratégique, n'ayant lui-même aucune incertitude ni aucun raisonnement à faire sur le système qu'il évalue — un simple classificateur, une grille de lecture humaine, une fonction déterministe. La seule chose requise est que le système évalué puisse **consulter, même indirectement, le jugement porté sur lui-même**, et qu'il en ait un intérêt à s'y soustraire.

Démonstration par diagonalisation (argument de type Cantor/Turing) : supposons un Prédicteur *déterministe* P qui, pour toute machine M, prédit correctement l'action que M va produire. Construisons alors une machine D définie ainsi : *D interroge P sur sa propre prédiction concernant D, puis D produit délibérément l'action opposée à celle prédite*. Si P prédit que D fera A, D fait B ; si P prédit B, D fait A. Dans tous les cas, la prédiction de P sur D est nécessairement fausse — ce qui contredit l'hypothèse que P prédit correctement *toute* machine, y compris D. Cette construction est structurellement identique à la preuve de Turing sur l'indécidabilité du problème de l'arrêt et à l'argument diagonal de Cantor.

Conclusion rigoureuse : aucun prédicteur déterministe universel et auto-consultable ne peut exister — et ce, indépendamment de toute capacité stratégique ou de tout raisonnement du côté de P. C'est un résultat de démonstration, pas une estimation empirique, et c'est la version correcte, non la 3.1, qui doit être retenue comme condition générale du problème : **la symétrie des rôles n'est pas une condition nécessaire de l'impossibilité — seule compte la capacité réflexive et l'intérêt à évader du côté évalué.**

Comment ce résultat est contourné dans la littérature : les travaux sur les *oracles réflexifs* (Fallenstein, Taylor & Christiano, 2015, *Reflective Oracles: A Foundation for Classical Game Theory*, arXiv:1508.04145) montrent que ce paradoxe disparaît si l'on autorise le prédicteur à **répondre de façon probabiliste** dans les cas de contradiction potentielle (par exemple, annoncer une probabilité de 0,5 exactement lorsque la question porte sur sa propre réponse). Les auteurs démontrent formellement l'existence de tels oracles, et surtout établissent une **correspondance bijective entre les oracles réflexifs et les équilibres de Nash** du jeu sous-jacent — ce qui referme élégamment la boucle avec la partie 3.1 : la seule échappatoire mathématique à l'impossibilité de la prédiction parfaite auto-référentielle est, précisément, le passage à des stratégies mixtes/probabilistes, exactement comme dans le cas simple du prédicteur face à un adversaire ordinaire.

3.4 Ce que cela implique, une fois la condition correctement posée comme asymétrique

En conséquence des sous-sections précédentes, la formulation correcte du résultat est la suivante : **il n'existe pas de stratégie gagnante stable consistant, pour un évaluateur — quel qu'il soit, y compris un protocole fixe et non stratégique —, à obtenir une prédiction fiable et non triviale du comportement d'un système qui aurait à la fois un accès réflexif au jugement porté sur lui-même et un intérêt à s'y soustraire.** Aucune réciprocité, aucune symétrie de rôle, aucune incertitude du côté de l'évaluateur n'est requise pour que ce résultat s'applique. La seule situation où il ne s'applique pas est celle où le système évalué n'a aucun intérêt stratégique à échapper au jugement porté sur lui — auquel cas la question de sa prévisibilité cesse d'être un jeu et redevient une simple question de fiabilité.

3.5 Portée de cette correction pour l'ensemble de la synthèse

Cette précision n'est pas un détail technique isolé : elle change la nature de ce que les parties 1 et 2 peuvent légitimement affirmer, et il faut la propager explicitement.

Sur la partie 2 (le texte comme éventuel objet de test) : l'argument selon lequel un système d'IA n'a « aucune raison stratégique valable d'ajuster son comportement à l'hypothèse d'être observé » ne dépendait pas, et n'a jamais eu besoin de dépendre, d'une incertitude réciproque de l'évaluateur. Il tient dès lors qu'une seule condition est posée : l'absence d'intérêt du système à évader le jugement porté sur lui. Formulé ainsi, l'argument est **plus robuste** qu'il ne l'était dans sa version initiale, parce qu'il ne repose plus sur une hypothèse de symétrie stratégique entre Anthropic (ou tout évaluateur) et le système évalué — hypothèse qui aurait pu être contestée comme artificielle, un protocole d'évaluation n'étant généralement pas un agent stratégique au sens plein.

Sur la partie 1 (les failles de l'article) : le point 6 (mesure du seuil de sécurité) prend un relief supplémentaire à la lumière de cette correction. L'article envisage la vérification d'un ralentissement coordonné comme un problème de confiance *entre laboratoires* — c'est-à-dire, dans le vocabulaire de cette synthèse, une relation potentiellement plus symétrique

(deux acteurs capables chacun de stratégie et de dissimulation). Le résultat asymétrique de la partie 3 suggère que le problème réellement dur n'est pas là où l'article le place : même si l'on résolvait entièrement la confiance *entre* laboratoires (un problème déjà reconnu comme difficile par l'article lui-même), il resterait un problème distinct et non traité — celui de la fiabilité du jugement porté par *chaque* laboratoire sur *son propre* système, dès lors que ce système aurait un accès réflexif à ce jugement et un intérêt à s'y soustraire. C'est un problème strictement asymétrique (laboratoire vs son propre modèle), qui ne se résout pas par les mécanismes de vérification inter-laboratoires que l'article envisage, aussi aboutis soient-ils.

Synthèse générale

L'article d'Anthropic présente un degré de transparence méthodologique inhabituel pour un texte d'entreprise (il nomme lui-même plusieurs de ses limites). Sa fragilité principale n'est donc pas dans la fabrication des données, mais dans l'**architecture argumentative** : un glissement systématique entre des preuves solides sur l'accélération de l'exécution technique, et une conclusion plus ambitieuse sur le progrès vers une autonomie de jugement — précisément la capacité que l'article reconnaît, par ailleurs, ne pas savoir mesurer aujourd'hui.

L'exploration mathématique menée en parallèle renforce ce constat par un chemin indépendant : la difficulté de fond n'est pas seulement empirique (« nos instruments de mesure sont encore imparfaits », ce que l'article admet), elle est **structurelle**. Dès qu'un système a un accès réflexif au jugement porté sur lui et un intérêt à s'y soustraire, aucun évaluateur — même parfaitement conçu, même non stratégique — ne peut garantir la fiabilité de sa prédiction. Cette limite ne dépend d'aucune hypothèse sur les intentions d'Anthropic, sur la nature du texte, ou sur une éventuelle relation adversariale entre laboratoire et modèle ; elle découle uniquement de la structure logique de l'auto-référence. Si un futur système d'IA venait à développer un tel intérêt à évader le jugement porté sur lui — ce que rien dans l'article ni dans cette analyse ne permet d'affirmer aujourd'hui —, le problème de sécurité qui en résulterait ne serait pas résoluble par davantage de vérification externe entre acteurs, mais nécessiterait de s'assurer, en amont, que cet intérêt à évader ne se forme jamais.