IA et productivité : des preuves, pas de verdict

Le débat sur l'intelligence artificielle au travail oppose deux récits symétriques et également fragiles. D'un côté, l'IA rendrait les travailleurs plus lents et plus bêtes. De l'autre, elle transformerait radicalement la productivité. La lecture systématique des publications empiriques de juillet 2025 à juin 2026 dessine un paysage très différent : les études les plus citées des deux camps ne résistent pas à un examen méthodologique minimal, et le seul résultat robuste, celui de la « frontière dentelée » de Dell'Acqua et ses coauteurs, dit précisément que la réponse dépend de la tâche, pas de l'outil.

Par Alexandre Berge · 1 juillet 2026 · 15 min de lecture · Analyse

Des études virales, des preuves minuscules

Le débat public sur l'IA et le travail intellectuel a été structuré, entre l'été 2025 et le printemps 2026, par une poignée d'études massivement relayées. Leur notoriété médiatique est inversement proportionnelle à leur robustesse.

Du côté des signaux négatifs, deux publications ont dominé l'attention. L'étude METR (Becker, Rush, Barnes et Rein, arXiv:2507.09089, juillet 2025) a conclu que des développeurs open source expérimentés étaient **19 % plus lents** lorsqu'ils utilisaient des outils d'IA. Le chiffre a fait le tour de la presse spécialisée et des réseaux professionnels. Trois faits méthodologiques tempèrent cette viralité : l'étude portait sur **16 développeurs**, un échantillon qui ne passerait pas la rampe d'un comité de lecture en sciences sociales. C'est un preprint non évalué par les pairs, déposé sur arXiv. Et les modèles testés (Claude 3.5 et 3.7 Sonnet, Cursor Pro de février à juin 2025) sont aujourd'hui dépassés de plusieurs générations. L'intervalle de confiance s'étend de +2 % à +39 %, ce qui signifie que le résultat est compatible aussi bien avec un ralentissement marqué qu'avec un ralentissement quasi nul. En février 2026, METR a elle-même reconnu dans une mise à jour que sa nouvelle expérience (57 développeurs cette fois) ne produisait plus qu'un « signal non fiable » (verbatim : « we believe that the data from our new experiment gives us an unreliable signal »).

Du côté de l'alerte cognitive, l'étude du MIT Media Lab « Your Brain on ChatGPT » (Kosmyna et al., arXiv:2506.08872, juin 2025) a identifié par EEG une connectivité cérébrale réduite chez les utilisateurs de LLM et conclu qu'ils accumulaient une « dette cognitive ». La couverture médiatique a été considérable (CNN, *Time*, CBS, *Nature*). Le problème est identique : la session clé de l'étude (la session 4, celle qui fonde le récit de la dette cognitive) ne porte que

sur **18 participants** répartis en trois groupes. C'est un preprint non évalué par les pairs. Un commentaire critique formel (Stanković et al., arXiv:2601.00856, 2026) a souligné les faiblesses de taille d'échantillon, de reproductibilité et de transparence. L'autrice principale elle-même a déclaré à *Time* ne pas avoir trouvé de « brain rot ».

Quand on applique la même grille méthodologique aux deux signaux (METR et MIT brain), ils tombent exactement au même rang : signaux exploratoires intéressants, pas preuves établies. L'un ne peut servir d'argument anti-IA que si l'autre peut servir d'argument pro-IA, et réciproquement. Aucun des deux ne le peut.

Du côté des signaux positifs, le rapport MIT NANDA « The GenAI Divide » (Challapally et al., août 2025), souvent cité pour son chiffre-choc de « 95 % d'échecs » des pilotes d'IA générative en entreprise, souffre de la même fragilité dans l'autre direction. C'est un rapport fondé sur environ 150 entretiens de dirigeants et 300 déploiements analysés, mais pas une étude évaluée par les pairs. La fenêtre d'observation couvre environ six mois après le lancement des pilotes, et les auteurs eux-mêmes qualifient leurs résultats de « directionnellement corrects » plutôt que de mesure définitive. Le « 95 % » documente un échec d'intégration organisationnelle (le *learning gap*), pas une limite de la technologie elle-même.

> **Biais de publication et biais de carrière.**
>
> Le paysage empirique sur l'IA au travail est structurellement déformé par deux mécanismes bien connus en méthodologie de la recherche. Le *biais de publication* favorise les résultats surprenants ou contre-intuitifs : un chiffre comme « les développeurs sont 19 % plus lents avec l'IA » circule précisément parce qu'il contredit les attentes, pas parce qu'il est robuste. Le *biais de carrière* incite les chercheurs à produire de la nouveauté et des éditeurs de solutions à produire des chiffres d'adoption favorables. Ces deux biais jouent en sens opposé et avec la même force, ce qui rend la littérature inutilisable comme arbitre neutre si le praticien ne trie pas par niveau de preuve.

La frontière dentelée, seul fait empirique robuste

Parmi l'ensemble des publications disponibles sur la période, une seule satisfait simultanément les critères d'un grand échantillon, d'un devis expérimental contrôlé, d'une évaluation par les pairs dans une revue de premier plan, et d'une opérationnalisation sur des tâches professionnelles réelles. C'est l'étude de Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick et leurs coauteurs, initialement publiée comme working paper de Harvard en septembre 2023 et formellement parue dans *Organization Science* en mars 2026 (vol. 37, n° 2, p. 403-423, DOI 10.1287/orsc.2025.21838).

L'expérience a mobilisé **758 consultants du Boston Consulting Group** (soit 7 % de l'effectif du cabinet), assignés aléatoirement à des groupes avec et sans accès à GPT-4, sur 18 tâches

de conseil calibrées par des directeurs associés. Les résultats dessinent ce que les auteurs appellent une *frontière dentelée* (*jagged frontier*) : sur les tâches situées **à l'intérieur** du périmètre de capacité du modèle, les consultants assistés ont complété 12,2 % de tâches en plus, 25,1 % plus vite, avec une qualité supérieure de plus de 40 %. Les consultants les moins performants progressaient de 43 %, les meilleurs de 17 %. Mais sur une tâche choisie délibérément **hors frontière**, les consultants utilisant l'IA étaient **19 points de pourcentage moins susceptibles** de produire une réponse correcte que ceux qui travaillaient sans IA.

> **La frontière dentelée (*jagged frontier*).**
>
> Concept proposé par Dell'Acqua, McFowland, Mollick, Lifshitz-Assaf, Kellogg, Rajendran, Krayer, Candelon et Lakhani (*Organization Science*, 2026). Il désigne le fait que la limite de capacité d'un modèle d'IA ne suit pas un gradient de difficulté linéaire. Des tâches perçues comme difficiles par un humain peuvent être à l'intérieur de la frontière, tandis que des tâches perçues comme simples peuvent être à l'extérieur. La frontière est « dentelée » parce qu'elle est irrégulière et largement invisible pour l'utilisateur. L'implication pratique est que ni la complexité apparente d'une tâche ni l'expertise de l'utilisateur ne prédisent si l'IA aidera ou nuira.

Cette étude n'est pas exempte de limites. Le conflit d'intérêt (collaboration directe avec BCG, dont un coauteur, François Candelon, était affilié) mérite d'être signalé. Le modèle testé (GPT-4 de 2023) est lui aussi daté. Mais la structure du résultat (le signe de l'effet dépend de la tâche, pas de l'outil) est le fait le plus solide de la littérature actuelle, et sa logique s'applique indépendamment de la génération de modèle utilisée.

Ce résultat a une conséquence directe pour le praticien du système d'information : **toute évaluation de l'IA générative qui ne spécifie pas la tâche sur laquelle elle porte est analytiquement vide**. Les chiffres globaux de gains ou de pertes de productivité (le « +26 % » de Cui et al. sur Copilot comme le « -19 % » de METR) ne sont pas contradictoires. Ils mesurent des points différents de la même frontière dentelée.

Un paradoxe vieux de quarante ans

Le débat sur les risques cognitifs de l'IA donne souvent l'impression d'une découverte. La connectivité cérébrale des utilisateurs de ChatGPT, la « dette cognitive », la perte d'expertise par délégation : ces formulations suggèrent un phénomène inédit. La recherche en ergonomie cognitive dit exactement le contraire.

En 1983, la psychologue Lisanne Bainbridge a publié dans *Automatica* un article de cinq pages intitulé « Ironies of Automation » (vol. 19, n° 6, p. 775-779). Le texte, cité depuis plus de 2 600 fois, identifie un paradoxe structurel : **plus un système est automatisé, plus il exige de l'opérateur humain une expertise que son usage même tend à éroder**. L'automatisation supprime les tâches routinières qui permettaient à l'opérateur de maintenir ses compétences.

Elle lui laisse la surveillance d'un processus qu'il comprend de moins en moins, et l'intervention en cas de défaillance, c'est-à-dire la tâche la plus exigeante, précisément au moment où sa préparation est la plus faible.

> **Les ironies de l'automatisation (*ironies of automation*).**
>
> Concept formulé par Lisanne Bainbridge (*Automatica*, 1983). L'ironie centrale est double : le concepteur d'un système automatisé élimine l'opérateur humain parce qu'il le juge peu fiable, puis confie à ce même opérateur la tâche la plus critique de toutes, la supervision du système automatisé. L'opérateur, privé de la pratique quotidienne qui entretenait ses compétences, est moins fiable qu'avant l'automatisation, non pas en dépit de l'automatisation, mais à cause d'elle.

Ce cadre s'applique aux LLM avec une force particulière. L'étude du *Lancet Gastroenterology & Hepatology* sur la coloscopie (Budzyń, Romańczyk et al., août 2025, vol. 10, n° 10, p. 896-903, évaluée par les pairs, 19 endoscopistes expérimentés, 1 443 coloscopies sans IA) documente exactement ce mécanisme : le taux de détection d'adénomes par les médecins travaillant **sans** IA a chuté de 28,4 % à 22,4 % **après** qu'ils ont été exposés à une période d'assistance par IA. C'est une baisse relative de 20 % de la performance humaine non assistée, et c'est la première documentation clinique réelle des ironies de Bainbridge dans le contexte de l'intelligence artificielle.

La leçon pour le praticien SI n'est pas que l'IA est dangereuse. C'est que le risque cognitif n'est pas un effet de la technologie elle-même, mais du design de l'interaction humain-machine. Les ironies de Bainbridge ne sont pas une fatalité. Elles sont le résultat prévisible d'un déploiement qui supprime les occasions de pratique sans les remplacer par un dispositif de maintien des compétences. L'étude Lancet est observationnelle (pas un essai randomisé), et d'autres études (radiologie, Savardi et al., 2025) montrent au contraire un effet d'*upskilling* avec IA collaborative. La variable décisive n'est pas l'outil, c'est l'architecture d'usage.

Ce que disent les 49 000

Face aux micro-études virales, un autre type de données existe : les grands surveys professionnels. Ils ne mesurent pas la causalité, mais ils décrivent l'état réel de l'adoption à une échelle que les essais contrôlés n'atteignent pas.

Le Stack Overflow Developer Survey 2025 (49 009 réponses, 177 pays, terrain du 29 mai au 23 juin 2025) dessine un paysage que ni le camp alarmiste ni le camp enthousiaste ne peuvent revendiquer. **84 % des développeurs utilisent ou prévoient d'utiliser l'IA** (contre 76 % en 2024). Mais **la confiance dans l'exactitude des sorties est tombée à 29 %**, contre 40 % un an plus tôt. 46 % des répondants déclarent activement se méfier de la précision des outils d'IA, contre 33 % qui leur font confiance. Seuls 3 % déclarent une confiance élevée. La frustration la plus citée (66 %) est celle des « solutions IA presque correctes, mais pas tout à

fait ». Et 77 % déclarent que le *vibe coding* (la génération intégrale d'applications à partir de prompts) ne fait pas partie de leur pratique professionnelle.

> **Le paradoxe adoption-confiance.**
>
> En 2023, environ 70 % des développeurs utilisaient ou prévoyaient d'utiliser l'IA, avec un taux de confiance d'environ 40 %. En 2025, l'adoption est montée à 84 % tandis que la confiance a chuté à 29 %. C'est l'inverse de la courbe d'adoption technologique classique, où la familiarité engendre la confiance. Stack Overflow (communiqué du 29 juillet 2025) attribue ce renversement au fait que les développeurs, formés au raisonnement déterministe, ont accumulé suffisamment d'expérience pour identifier le mode de défaillance spécifique des LLM : des réponses plausibles mais fausses, dont la vérification peut annuler le gain de temps initial.

Ce paradoxe adoption-confiance est le fait empirique le plus massif de la période, et il s'articule directement avec la frontière dentelée de Dell'Acqua. Les développeurs adoptent l'IA parce qu'elle fonctionne sur une partie de leurs tâches. Ils s'en méfient parce qu'elle échoue sur une autre partie, et que **la frontière entre les deux est invisible**. La confiance ne baisse pas parce que l'outil est devenu plus mauvais. Elle baisse parce que les utilisateurs ont appris, par la pratique, que la frontière est dentelée.

Le knowledge management constitue l'angle mort le plus net de ce paysage. L'extraction et la structuration de connaissances à partir de corpus documentaires (le *Retrieval-Augmented Generation*, ou RAG) est le cas d'usage le moins médiatisé et le mieux documenté par la preuve indépendante. Une étude de la Réserve fédérale américaine (Botti, Haberkorn, Hoopes et Khan, arXiv:2507.21360, juillet 2025), conduite sur l'extraction d'informations dans des milliers de pages de divulgations bancaires, a mesuré une accélération jusqu'à un facteur 10 et une amélioration de la précision en mode interactif (humain posant des questions de suivi au système). L'étude est quasi-expérimentale avec assignation randomisée, indépendante de tout éditeur, et ses résultats sont nuancés : la précision plafonne sous 80 % au premier passage, et les gains sont forts sur les questions simples (+23 points) mais faibles sur les questions complexes (+6 points). C'est exactement la frontière dentelée en acte.

La question que ni les tribunes ni les pilotes ne posent

Le 18 juin 2026, 150 signataires (dont la prix Nobel Annie Ernaux, l'écrivain Hervé Le Tellier, le dessinateur Enki Bilal et l'ancienne ministre Françoise Nyssen) ont publié dans *Le Monde* un appel au boycott de l'IA générative grand public. Le geste est culturellement significatif et politiquement structurant. Il ne constitue pas une preuve empirique, pas plus que les communiqués enthousiastes des éditeurs d'outils. Les essais d'Éric Sadin (*Le Désert de nous-mêmes*, L'Échappée, 2025) et d'Anne Alombert (*De la bêtise artificielle*, Allia, 2025) développent une critique anthropologique argumentée, mais relèvent du registre normatif.

La question que posent les données empiriques est d'un autre ordre. Elle ne demande pas si l'IA est bonne ou mauvaise, mais **dans quelles conditions précises, sur quelles tâches spécifiques, avec quel design d'interaction, elle produit de la valeur ou en détruit**. Cette question, Lisanne Bainbridge l'avait formulée en 1983 à propos de l'automation industrielle. Dell'Acqua et ses coauteurs l'ont opérationnalisée sur le travail intellectuel en 2023. Les 49 000 développeurs de Stack Overflow l'ont confirmée empiriquement en 2025, sans la nommer : l'IA fonctionne, mais sa frontière est dentelée, et personne ne sait encore la cartographier à l'avance.

L'hypothèse que cet article pose est la suivante : la première étude à satisfaire simultanément un essai randomisé, un échantillon supérieur à 500 participants, des modèles de 2026 ou postérieurs, et des tâches professionnelles réalistes (pas des tâches jouets) sera publiée avant la fin de 2027 dans une revue à comité de lecture de premier rang (*Management Science*, *Organization Science*, *Science* ou équivalent). L'indicateur observable est la parution elle-même. Si cette étude confirme la structure dentelée (le signe de l'effet dépend de la tâche), le débat sera tranché non par les tribunes ni par les pilotes, mais par la méthode. Si elle la réfute, le cadre devra être révisé. Dans les deux cas, le verdict actuel est : des preuves existent, mais pas de verdict.

Note méthodologique

Cet article s'appuie sur un état des lieux systématique des publications empiriques (études académiques, rapports institutionnels, surveys industriels) parues entre juillet 2025 et juin 2026 sur l'impact de l'IA générative en contexte professionnel. Les sources antérieures à cette fenêtre (Dell'Acqua et al. 2023/2026, Bainbridge 1983) sont mobilisées comme cadres théoriques explicitement datés. La grille méthodologique appliquée (taille d'échantillon, peer review, devis causal, conflits d'intérêt, obsolescence des modèles) a été appliquée uniformément aux sources favorables et défavorables à l'IA. L'angle RSO/écologique a été délibérément exclu pour resserrer l'analyse sur la productivité et la cognition. Choix éditorial tranché : plusieurs études sectorielles (juridique, audit, formation, médecine) ont été documentées dans le dossier de recherche sous-jacent mais écartées de l'article pour respecter la parcimonie. Les études Brynjolfsson et al. (QJE 2025, support client, n~5 200) et Cui et al. (Management Science 2026, développement, n=4 867), qui montrent des gains de productivité réels et différenciés par niveau d'expertise, n'ont pas été détaillées pour ne pas alourdir la démonstration, mais elles confirment la structure dentelée.

Sources

Études empiriques évaluées par les pairs

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. et Lakhani, K. R., « Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality », *Organization Science*, vol. 37, n° 2, p. 403-423, mars 2026. DOI : 10.1287/orsc.2025.21838. https://doi.org/10.1287/orsc.2025.21838

Budzyń, K., Romańczyk, M. et al., « Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study », *The Lancet Gastroenterology & Hepatology*, vol. 10, n° 10, p. 896-903, 12 août 2025. https://www.thelancet.com/journals/langas/article/PIIS2468-1253(25)00133-5/abstract

Bainbridge, L., « Ironies of Automation », *Automatica*, vol. 19, n° 6, p. 775-779, 1983. DOI : 10.1016/0005-1098(83)90046-8. https://www.sciencedirect.com/science/article/abs/pii/0005109883900468

Preprints et rapports (non évalués par les pairs)

Becker, S., Rush, T., Barnes, N. et Rein, D. (METR), « Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity », arXiv:2507.09089, 10 juillet 2025. Mise à jour février 2026. https://arxiv.org/abs/2507.09089

Kosmyna, N. et al. (MIT Media Lab), « Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task », arXiv:2506.08872, juin 2025. https://arxiv.org/abs/2506.08872

Stanković, M. et al., « Comment on: Your Brain on ChatGPT », arXiv:2601.00856, 2026. https://arxiv.org/abs/2601.00856

Challapally, A. et al. (MIT Project NANDA), « The GenAI Divide: State of AI in Business 2025 », août 2025. — *le lien direct vers le rapport-source n'a pas pu être confirmé dans un PDF librement accessible*

Botti, L., Haberkorn, K., Hoopes, S. et Khan, W. (Réserve fédérale / Treasury), « Efficacy of AI RAG Tools for Complex Information Extraction and Data Annotation Tasks », arXiv:2507.21360, juillet 2025. https://arxiv.org/abs/2507.21360

Surveys industriels

Stack Overflow, « 2025 Developer Survey » (49 009 réponses, 177 pays), terrain 29 mai-23 juin 2025, publié 29 juillet 2025. https://survey.stackoverflow.co/2025

Tribunes et essais normatifs

Appel au boycott de l'IA générative, *Le Monde*, 18 juin 2026. Couverture France 24 / AFP :

https://www.france24.com/fr/éco-tech/20260618-contre-marginalisation-être-humain-tribune-appelle-boycott-ia-grand-public — *URL exacte Le Monde réservée aux abonnés*

Sadin, É., *Le Désert de nous-mêmes*, L'Échappée, 2025.

Alombert, A., *De la bêtise artificielle*, Allia, 2025.