L’attention et la charge cognitive du travail assisté

Étude 14 — Réagencer le travail à l’âge des agents

Mathieu Guglielmino

2026-06-25

Le fil

ℭ Journal des mises à jour

- **2026-07-07** — Section 5 : la bascule exécution→supervision mesurée sur 319 travailleurs du savoir (Lee et al., CHI 2025) — l’effort baisse le moins sur l’évaluation (55 %) ; fig05.
- **2026-06-25** — Création du dossier et du journal (étude 14, pilier Design d’interaction) : thèse de la bascule exécution → supervision, 4 schémas, ancrage DARES/Microsoft/Mark, module 14 du baromètre cadres-rh étoffé à 5 items.

La thèse

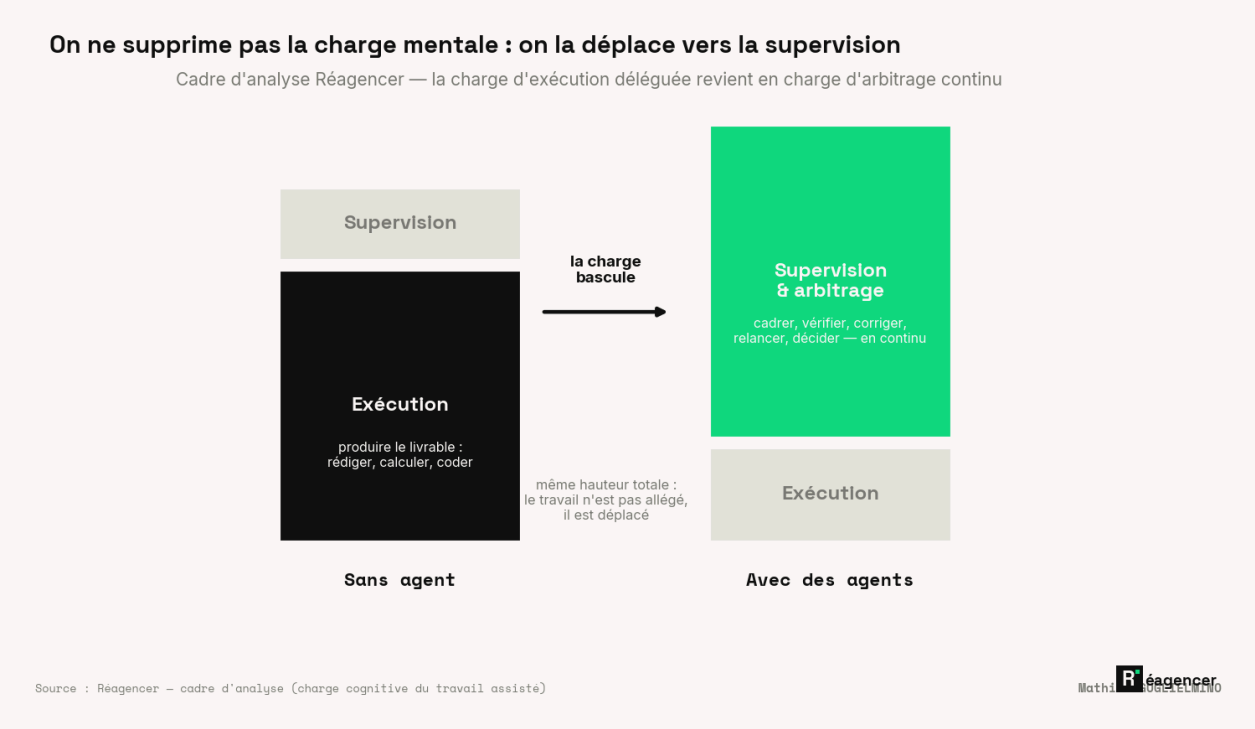
On n’a pas supprimé la charge mentale du travail : on l’a déplacée. La charge d’exécution — produire le livrable, rédiger, calculer, coder — est visible, bornée, reconnue ; on la délègue à la machine. Elle revient sous une autre forme : cadrer, vérifier, corriger, relancer, décider. Une charge de **supervision et d’arbitrage** — diffuse, continue, peu visible. Le risque du travail assisté n’est donc pas la surcharge de tâches mais l’**épuisement par sur-sollicitation** : on travaille entouré d’agents qui proposent, interrompent et relancent sans fin. Corollaire Réagencer : **la charge de supervision est le coût caché du travail assisté — réel, mais non mesuré, donc non protégé.** Ce dossier n’est pas un réquisitoire contre l’IA. La fragmentation de l’attention précède l’agent de trente ans ; l’agent ne crée pas le problème, il se déverse sur une attention déjà saturée et y ajoute une couche de vigilance. La question n’est pas « pour ou contre » — c’est « à quelles conditions de design ».

---

1. La charge ne disparaît pas, elle bascule vers la supervision

Le premier réflexe, devant un agent qui produit à notre place, c'est de croire qu'on s'est allégé. On a délégué l'exécution : le rapport se rédige seul, le code s'écrit seul, le tableau se remplit seul. Mais le travail n'a pas diminué — il a changé de nature.

```
analyse.fig_bascule(d);
```



C'est un cadre d'analyse, pas une mesure. Le schéma ci-dessus ne pèse pas deux charges sur le même instrument : il pose une **grille de lecture** Réagencer, étayée par les chiffres des sections suivantes. Les hauteurs sont illustratives — elles disent un report, pas une quantité mesurée. Aucune enquête française ne mesure encore la « charge de supervision » propre aux agents (cf. dernière section).

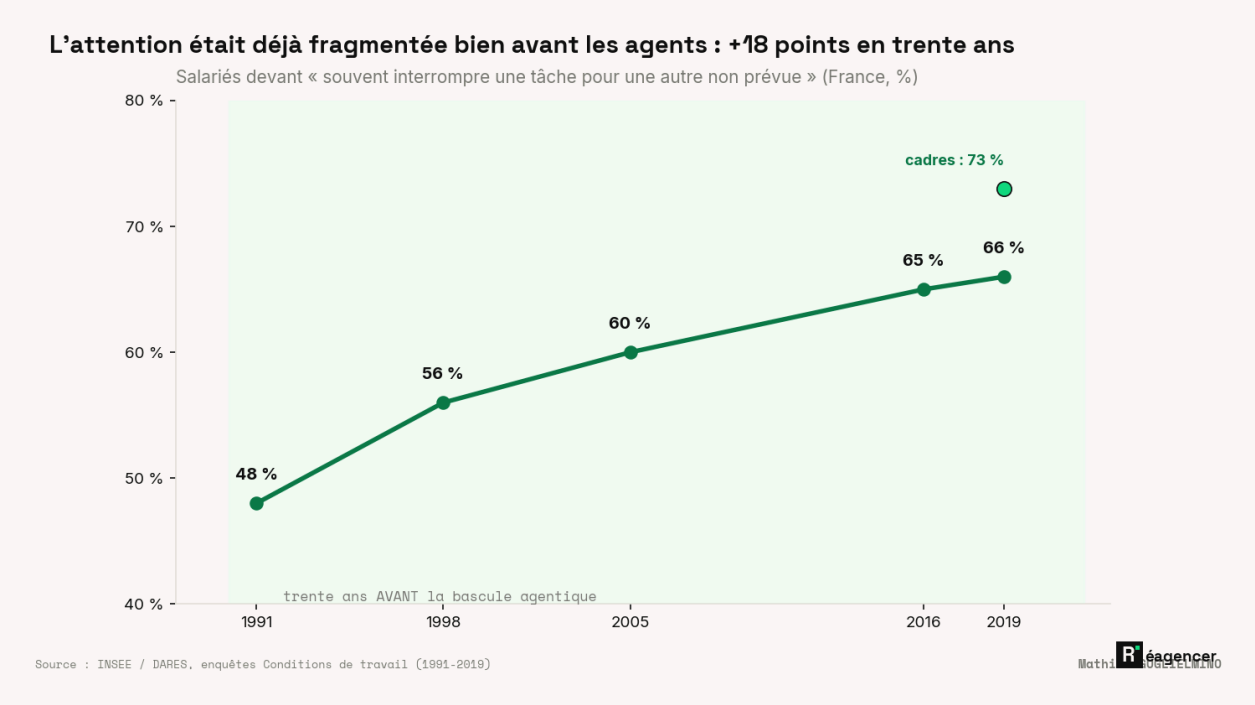
L'ergonomie sait depuis longtemps qu'une activité se décompose en trois plans : le travail **prescrit** (ce qu'on demande de faire), le travail **réel** (ce qu'il faut effectivement faire pour y arriver) et la charge **vécue** (ce que ça coûte à la personne). L'agent agit sur le prescrit : la tâche d'exécution sort de la fiche. Mais le réel, lui, se reconstitue ailleurs. Il faut **cadrer** la demande pour que la machine produise quelque chose d'utilisable, **vérifier** ce qu'elle rend, **corriger** ce qui ne va pas, **relancer** quand c'est à côté, et **décider** si la sortie part en l'état ou non. Aucune de ces opérations n'apparaît dans une fiche de poste. Elles sont continues, peu visibles, et exigent un effort cognitif d'un genre précis : rester en alerte sur la qualité d'un travail qu'on n'a pas fait soi-même.

C'est le déplacement central de ce dossier. La charge d'exécution était bornée : on commençait, on finissait, on reconnaissait l'effort. La charge de supervision n'a pas de fin nette — elle suit le rythme de la machine, qui ne se fatigue pas. On ne travaille plus moins ; on travaille autrement, sur un registre — la vigilance — qui s'use d'une autre manière. Le reste de l'étude pose les chiffres qui rendent ce cadre crédible : d'abord que le terrain attentionnel était déjà abîmé avant l'agent, ensuite que l'agent s'y ajoute, enfin que le ressenti de soulagement masque un coût qui n'a fait que se déplacer.

2. L’attention était déjà fragmentée bien avant les agents

Avant d’imputer quoi que ce soit aux agents, il faut regarder l’état du terrain. Et le terrain était déjà fissuré.

```
analyse.fig_fragmentation(d);
```



La part de salariés français déclarant devoir « souvent interrompre une tâche pour une autre non prévue » n’a fait que monter sur trente ans : **48 %** en 1991, **56 %** en 1998, **60 %** en 2005, **65 %** en 2016, **66 %** en 2019[1]. Soit **+18 points** avant même que le premier agent génératif n’arrive sur un poste de travail. Les cadres — la population la plus exposée à l’assistance logicielle — sont à **73 %** en 2019, nettement au-dessus de la moyenne.

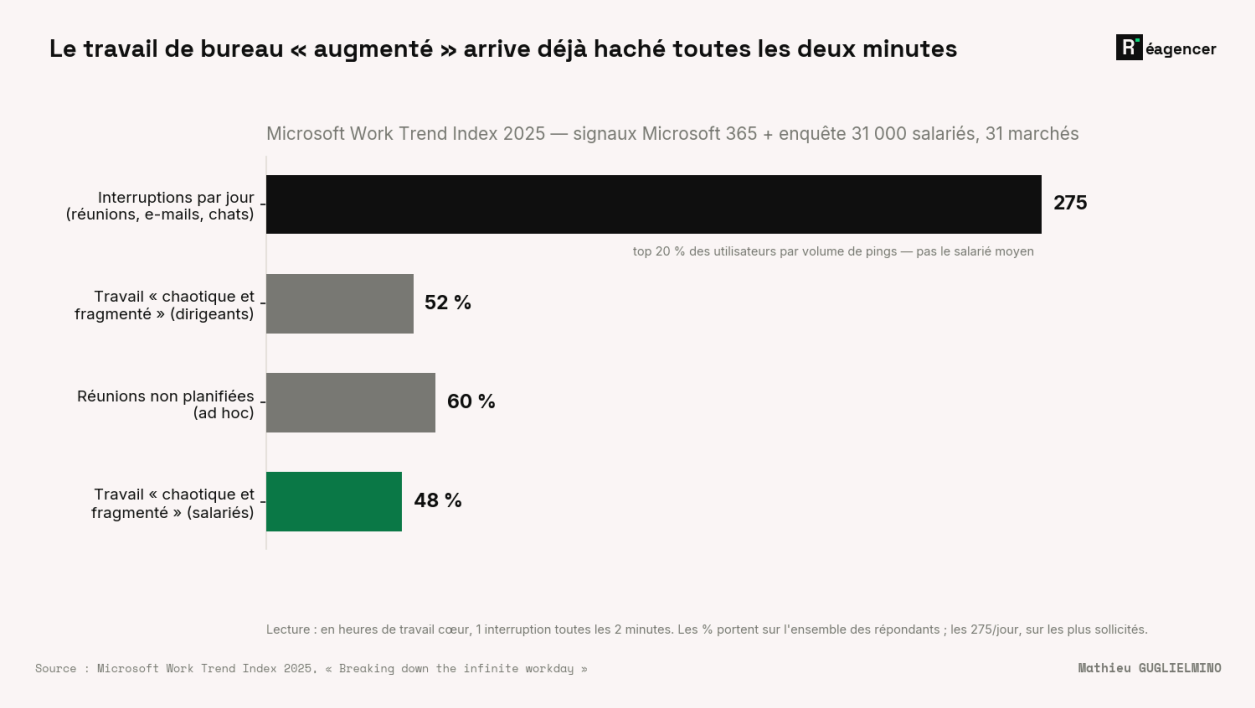
L’enseignement est net, et il vaut garde-fou pour tout le reste. L’agent n’invente pas l’attention hachée : il débarque sur une attention déjà saturée. Quiconque a vu un poste de travail de bureau ces dix dernières années le sait sans enquête — la journée se passe à être interrompu, par un message, une notification, une demande latérale, une réunion qui s’invite. Ce qui change avec l’agent, ce n’est pas l’apparition de l’interruption : c’est l’ajout d’une source d’interruption de plus, et d’une nature nouvelle — une machine qui propose, signale, relance, en continu, et qu’il faut surveiller. On part d’un seuil déjà élevé. C’est la raison pour laquelle le sujet mérite mieux qu’un slogan : le problème est ancien, mesuré, et l’agent n’en est qu’un accélérateur.

[1] INSEE / DARES (Bué, Coutrot, Guignon), « L’évolution des conditions de travail », série « interruptions fréquentes » (devoir souvent interrompre une tâche pour une autre non prévue) : 1991 = 48 %, 1998 = 56 %, 2005 = 60 %. https://www.insee.fr/fr/statistiques/fichier/1374396/EMPLOIR08g.PDF — prolongée par DARES, « Quelles étaient les conditions de travail en 2019, avant la crise sanitaire » : 2016 = 65 %, 2019 = 66 % (cadres 73 %). https://dares.travail-emploi.gouv.fr/publication/quelles-etaient-les-conditions-de-travail-en-2019-avant-la-crise-sanitaire

3. Le poste de travail « augmenté » arrive déjà haché toutes les deux minutes

Si la fragmentation est ancienne côté français, l'instrumentation logicielle du travail de bureau en donne une mesure plus crue, et plus récente.

```
analyse.fig_cadence(d);
```



Le Work Trend Index 2025 de Microsoft — construit sur des signaux d'usage de Microsoft 365 et une enquête de 31 000 salariés dans 31 marchés — compte **275 interruptions par jour** (réunions, e-mails, messages instantanés), soit une rupture environ **toutes les 2 minutes** en heures de travail cœur[2]. Dans la même enquête, **48 %** des salariés et **52 %** des dirigeants qualifient leur travail de « chaotique et fragmenté », et **60 %** des réunions sont non planifiées, ad hoc.

Le « 275 par jour » n'est pas le salarié moyen. C'est le **top 20 %** des utilisateurs par volume de pings — les plus sollicités, pas la médiane (méthodologie Microsoft). Le chiffre dit l'intensité du haut de la distribution, pas la norme. Les pourcentages « chaotique et fragmenté » et « réunions ad hoc », eux, portent sur l'ensemble des répondants. On ne mélange pas les deux échelles.

Même en retenant le caveat, la lecture tient : le terrain logiciel sur lequel l'agent vient se poser est déjà sur-sollicité. Quand on greffe un assistant — qui suggère pendant la frappe, qui signale une révision, qui propose une action — sur un poste où la concentration est déjà rompue toutes les deux minutes pour les plus exposés, on n'ajoute pas une aide dans le calme : on ajoute une voix de plus dans un environnement déjà bruyant. La question de design est là, exactement : un agent peut soit absorber une partie de ce bruit (en regroupant, en filtrant, en attendant le bon moment), soit en rajouter. Rien ne garantit, par défaut, qu'il fasse le premier.

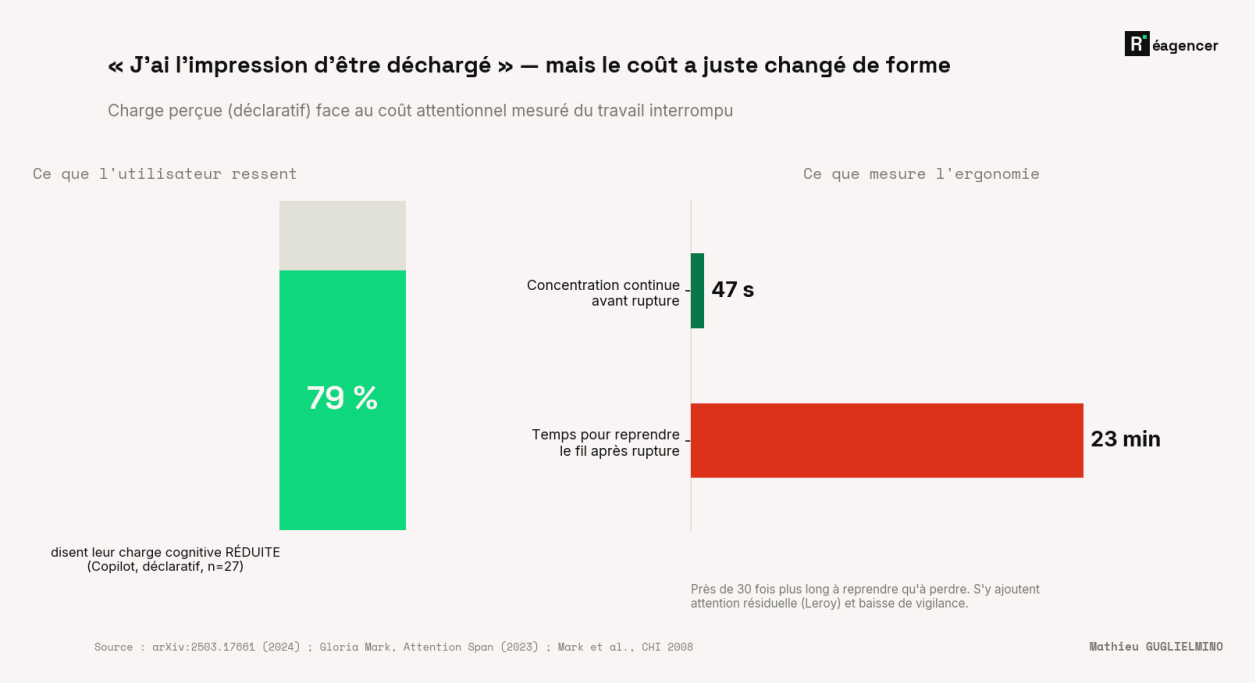
---

[2] Microsoft, Work Trend Index 2025 — « Breaking down the infinite workday » (signaux Microsoft 365 + enquête de 31 000 salariés dans 31 marchés). Chiffres cités : 275 interruptions/jour (top 20 % des utilisateurs par volume de pings, pas la moyenne), 1 interruption toutes les 2 minutes en heures cœur, 48 % des salariés / 52 % des dirigeants « travail chaotique et fragmenté », 60 % de réunions ad hoc. https://www.microsoft.com/en-us/worklab/work-trend-index/breaking-down-infinite-workday

4. « J'ai l'impression d'être déchargé » — mais le coût a juste changé de forme

Voici le cœur de l'honnêteté de ce dossier, et son point de doute. Car les utilisateurs, eux, ne se plaignent pas. Ils disent l'inverse.

```
analyse.fig_paradoxe(d);
```



Dans une étude qualitative sur l'assistant Copilot, **79 %** des utilisateurs déclarent leur charge cognitive réduite[3]. Le ressenti est massif, et il est sincère. Mais il faut le tenir à distance de ce qu'il est.

Le « 79 % » est un déclaratif, pas une mesure ergonomique — et n = 27. C'est un ressenti recueilli sur un très petit échantillon qualitatif, anglo-saxon. « Je me sens déchargé » est une information précieuse sur l'expérience vécue ; ce n'est pas une mesure de la charge réelle. Les deux peuvent diverger — c'est tout le sujet.

Car ce que mesure l'ergonomie, côté attention, raconte une autre histoire. La durée de concentration continue sur un écran avant de basculer ailleurs s'est effondrée : **150 secondes** en 2004, 75 en 2012, **47 secondes** en 2023[4]. Or il faut en moyenne **23 minutes** (23 min 15 s précisément, dans l'article fondateur) pour reprendre le fil d'une tâche après une interruption[5]. L'asymétrie est brutale : on perd le fil en moins d'une minute, on met près de **trente fois plus longtemps** à le retrouver. S'y ajoute l'**attention résiduelle** décrite par Sophie Leroy : après un changement de

[3] « A Qualitative Study of User Perception of M365 AI Copilot », arXiv:2503.17661, 2024 — 79 % des utilisateurs déclarent leur charge cognitive réduite (étude qualitative déclarative, n = 27). https://arxiv.org/abs/2503.17661

[4] Gloria Mark, Attention Span (2023) — durée de concentration continue par écran : 150 s (2004) → 75 s (2012) → 47 s (2023). https://gloriamark.com/attention-span/ — et Mark, Gudith & Klocke, « The Cost of Interrupted Work: More Speed and Stress », CHI 2008 : 23 min 15 s pour reprendre le fil après interruption. https://doi.org/10.1145/1357054.1357072

[5] Gloria Mark, Attention Span (2023) — durée de concentration continue par écran : 150 s (2004) → 75 s (2012) → 47 s (2023). https://gloriamark.com/attention-span/ — et Mark, Gudith & Klocke, « The Cost of Interrupted Work: More Speed and Stress », CHI 2008 : 23 min 15 s pour reprendre le fil après interruption. https://doi.org/10.1145/1357054.1357072

tâche, une part du traitement cognitif reste accrochée à la tâche précédente — on n'est jamais tout à fait revenu[6].

Et c'est précisément là que le travail assisté creuse un écart entre ressenti et coût réel. À mesure que l'agent devient capable, l'humain relâche son effort propre sur la production — c'est rationnel, c'est même le but. Mais le coût ne disparaît pas : il migre vers la **détection d'erreur**. Trois mécanismes documentés se conjuguent. La complaisance d'automatisation : on fait confiance à une sortie d'autant plus qu'elle a souvent été bonne, et on baisse la garde. La baisse de vigilance : surveiller en continu une sortie qu'on ne produit pas est une tâche attentionnelle coûteuse, qui se dégrade avec le temps. La fatigue de relecture : vérifier un livrable qu'on n'a pas écrit demande un effort différent, et l'œil glisse. Le résultat est un paradoxe net : on se sent déchargé parce que l'exécution est partie ; on est en réalité chargé d'une vigilance diffuse, plus difficile à nommer, donc plus difficile à protéger. « Je me sens déchargé » n'est pas « la charge a disparu ». Elle a changé de forme, et s'est rendue invisible.

---

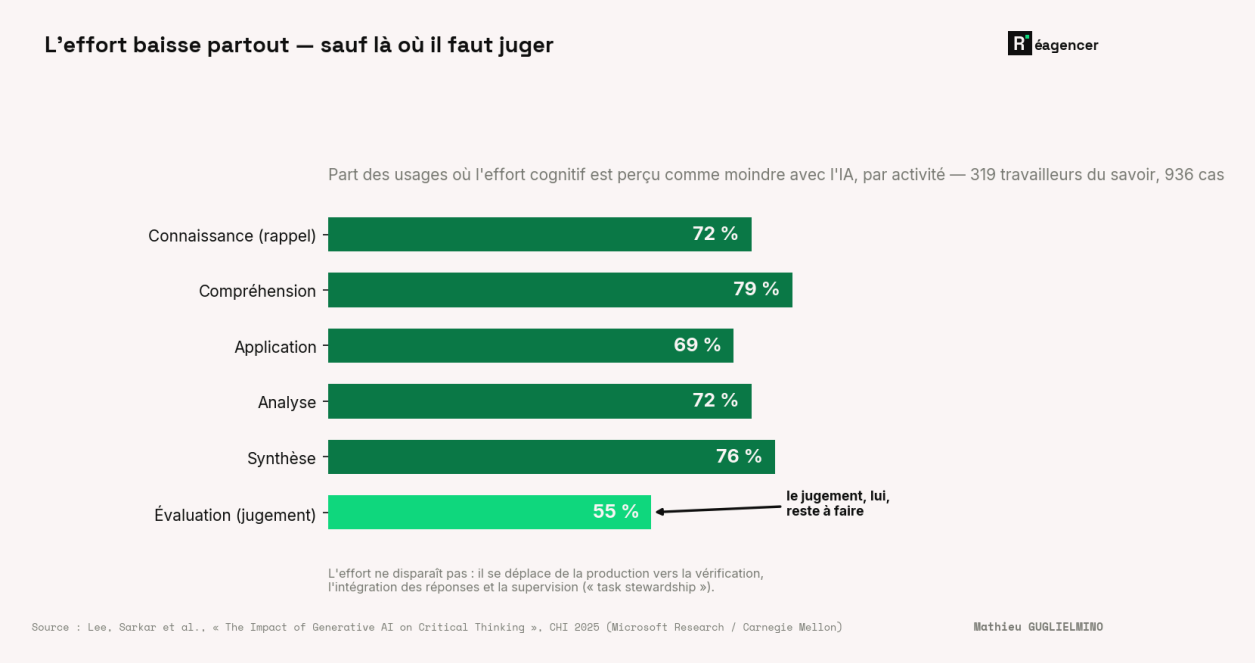
5. La bascule, enfin mesurée : tout s'allège, sauf le jugement

- Jusqu'ici, la bascule exécution → supervision n'était qu'un **cadre** — une grille de lecture étayée par des chiffres épars (fragmentation ancienne, poste saturé, paradoxe du ressenti) mais jamais pesée sur le geste même qu'elle prétend décrire. Une étude CHI 2025 de Microsoft Research et Carnegie Mellon change ce statut : elle mesure le déplacement, sur le travail réel de travailleurs du savoir[7].

```
analyse.fig_effort_bloom(d);
```

[6] Sophie Leroy, « Why Is It So Hard to Do My Work? The Challenge of Attention Residue When Switching Between Work Tasks », Organizational Behavior and Human Decision Processes 109(2), 2009 — concept d'attention residue : après un changement de tâche, une part du traitement cognitif reste accrochée à la tâche précédente. https://doi.org/10.1016/j.obhdp.2009.04.002

[7] Hao-Ping (Hank) Lee, Advait Sarkar et al., « The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers », CHI 2025 (Microsoft Research / Carnegie Mellon University) — n = 319 travailleurs du savoir, 936 cas d'usage réels analysés. Part des cas où l'effort cognitif perçu est « moindre » avec l'IA générative, par niveau de la taxonomie de Bloom : Connaissance/rappel 72 %, Compréhension 79 %, Application 69 %, Analyse 72 %, Synthèse 76 %, Évaluation 55 %. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/



- Lee, Sarkar et al. ont fait analyser **936 cas d'usage réels** de l'IA générative par **319 travailleurs du savoir**, puis classé chaque cas selon le niveau de la taxonomie de Bloom qu'il mobilisait — de la simple restitution d'information au jugement. Résultat : l'effort cognitif perçu baisse presque partout, et fort. Connaissance/rappel **72 %**, Compréhension **79 %**, Application **69 %**, Analyse **72 %**, Synthèse **76 %** des cas jugés « moins » ou « beaucoup moins » coûteux avec l'IA. C'est le « je me sens déchargé » de la section précédente, mais cette fois quantifié sur près d'un millier de cas d'usage, pas sur une poignée d'entretiens.
- Un seul niveau résiste : l'**évaluation**, c'est-à-dire le jugement — arbitrer, trancher, juger de la valeur d'une sortie. Là, l'effort perçu ne baisse que dans **55 %** des cas, le score le plus bas des six, loin derrière la Compréhension (79 %). Et les auteurs ne se contentent pas de le constater : ils nomment le déplacement dans les termes mêmes du cadre Réagencer. L'effort ne disparaît pas, il **se déplace** — de la collecte d'information vers la **vérification**, de la résolution de problème vers l'**intégration** de la réponse de l'IA, de l'exécution de la tâche vers la **supervision de la tâche** (« task stewardship »). C'est ici qu'une donnée externe, indépendante, vient confirmer le vocabulaire que ce dossier utilisait par hypothèse depuis le départ.
- Le second résultat de l'étude est le plus retors, et il concerne directement le pilier Design d'interaction. Une plus grande **confiance dans l'IA** est associée à **moins** d'esprit critique exercé ; une plus grande **confiance en soi** est associée à **plus** d'esprit critique. Autrement dit, un agent qui inspire une confiance aveugle n'allège pas seulement l'exécution : il érode la vigilance même dont dépend la supervision — le registre précisément qu'il faudrait renforcer si l'effort s'y déplace. Ce n'est pas un problème de puissance du modèle, c'est un problème de conception de l'interaction : comment un agent doit-il se présenter pour ne pas endormir le jugement qu'on lui demande, en creux, d'exercer sur lui ?
- **Lee et al. quantifie et nomme le déplacement — il ne le referme pas.** L'étude reste **déclarative** (l'effort perçu, pas une mesure comportementale du temps ou de la charge réelle) et **anglo-saxonne** (échantillon de travailleurs du savoir américains). Elle confirme, à grande échelle et avec le bon vocabulaire, que l'effort se déplace vers la vérification et la supervision — mais elle ne le mesure ni en France, ni sur un terrain observé. C'est exactement ce que le **module 14 du baromètre cadres-rh** (comportemental, français) doit encore produire.

Ce que la donnée ne dit pas encore

- **Angle mort majeur — la charge de supervision n'est pas mesurée en France, ni sur du comportemental.** Ce dossier établit trois choses sur données ouvertes : la fragmentation de l'attention est ancienne, mesurée et en hausse en France (DARES) ; le poste de travail logiciel est déjà saturé d'interruptions (Microsoft) ; le coût attentionnel d'une rupture est documenté (Mark, Leroy). Lee et al. ajoute une quatrième pierre : le déplacement de l'effort vers la vérification, l'intégration et la supervision est désormais mesuré et nommé, sur 319 travailleurs du savoir[8] — mais en restant déclaratif et anglo-saxon. Aucune enquête française ne mesure encore ce qui est au centre de la thèse : le **report de la charge d'exécution vers l'arbitrage** sur le terrain français, le nombre de validations par jour imputables à un agent, le sentiment de **contrôle vs débordement** face aux sorties d'IA, la fatigue de supervision elle-même. Les seules données « charge réduite » disponibles restent déclaratives et anglo-saxonnes. C'est exactement ce que le **module 14 du baromètre cadres-rh** est conçu pour produire — cinq items (14.1 à 14.5) : ressenti de charge, interruptions/jour imputables aux agents, sentiment de contrôle vs débordement, fatigue de supervision, et une question miroir côté socle (la délégation augmente-t-elle le sentiment de maîtrise ou de débordement ?). Le chiffre fondateur visé : la part de cadres qui se sentent déchargés de l'exécution mais débordés par la supervision.

Ce qu'il faut en retenir

Le travail assisté ne supprime pas la charge mentale : il la déplace de l'exécution vers la supervision. La charge d'exécution était visible, bornée, reconnue ; la charge de supervision est diffuse, continue, peu visible — et c'est précisément parce qu'elle est invisible qu'elle n'est ni mesurée, ni reconnue, ni protégée. C'est le coût caché du travail assisté.

- Quatre chiffres tiennent désormais ce cadre. La fragmentation de l'attention précède l'agent de trente ans (+18 points entre 1991 et 2019) : l'agent n'invente pas le problème, il s'ajoute à une attention déjà saturée. Le poste de travail de bureau est déjà rompu toutes les deux minutes pour les plus sollicités : l'agent débarque dans un environnement bruyant, pas dans le calme. Le ressenti de soulagement (79 % se disent déchargés) coexiste avec un coût attentionnel mesuré qui, lui, n'a pas disparu — il a migré vers la vigilance et la détection d'erreur, là où l'œil glisse et la garde baisse. Et sur 936 cas d'usage réels, l'effort baisse partout sauf sur le jugement (55 %) — la mesure, enfin, du déplacement que ce dossier posait en cadre[9].

La nuance praticien doit rester jusqu'au bout : ce n'est pas un procès de l'IA. La fragmentation est antérieure ; l'agent l'accélère mais ne l'a pas créée. Le vrai sujet, celui du pilier Design d'interaction, n'est pas « faut-il des agents » mais « **quel design d'agent réduit la charge de supervision au lieu de l'augmenter** ». Un agent qui interrompt à chaque suggestion, qui exige

[8] Hao-Ping (Hank) Lee, Advait Sarkar et al., « The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers », CHI 2025 (Microsoft Research / Carnegie Mellon University) — n = 319 travailleurs du savoir, 936 cas d'usage réels analysés. Part des cas où l'effort cognitif perçu est « moindre » avec l'IA générative, par niveau de la taxonomie de Bloom : Connaissance/rappel 72 %, Compréhension 79 %, Application 69 %, Analyse 72 %, Synthèse 76 %, Évaluation 55 %. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

[9] Hao-Ping (Hank) Lee, Advait Sarkar et al., « The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers », CHI 2025 (Microsoft Research / Carnegie Mellon University) — n = 319 travailleurs du savoir, 936 cas d'usage réels analysés. Part des cas où l'effort cognitif perçu est « moindre » avec l'IA générative, par niveau de la taxonomie de Bloom : Connaissance/rappel 72 %, Compréhension 79 %, Application 69 %, Analyse 72 %, Synthèse 76 %, Évaluation 55 %. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

une validation pour chaque pas, qui propose sans hiérarchiser, déverse sa propre charge sur l'humain. Un agent qui regroupe, qui attend le bon moment, qui sait dire ce dont il n'est pas sûr et qui demande un arbitrage seulement quand il compte — celui-là allège vraiment. La différence n'est pas dans le modèle : elle est dans la manière dont la machine partage la vigilance avec la personne. C'est une question de conception, pas de puissance. Et tant qu'on ne mesure pas la charge de supervision, on conçoit à l'aveugle. Le module 14 du baromètre cadres-rh est là pour ouvrir l'œil.

Sources et références

Calculs et schémas : Réagencer. Journal de recherche en cours — Réagencer, juin 2026. Données ouvertes au 2026-06-25.