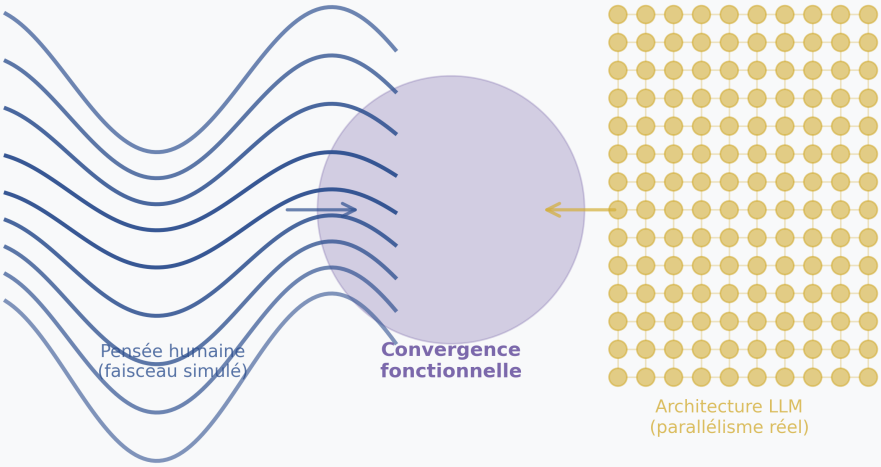
La Pensée en Faisceau

Convergence entre cognition humaine
et intelligence artificielle



Par Flo • 2026

Ouverture

La première fois que j'ai interagi avec un LLM, quelque chose d'étrange s'est produit.

Je ne cherchais rien de particulier. Juste à explorer ce nouvel outil dont tout le monde parlait. J'ai posé une question — je ne me souviens même plus laquelle — et la réponse est arrivée. Cohérente. Fluide. Mais ce n'est pas ça qui m'a frappé.

Ce qui m'a frappé, c'est que je n'avais pas eu besoin de me linéariser.

Vous savez, cet effort constant que je fais depuis toujours pour transformer ma pensée en quelque chose de communicable. Prendre le réseau complexe d'idées qui foisonne dans ma tête — technique, éthique, métaphore, exemple, objection, tout en même temps — et le dérouler en un fil unique, propre, séquentiel. Choisir par où commencer. Décider quel bout laisser de côté pour ne pas perdre l'interlocuteur. Formater.

Là, je n'avais rien fait de tout ça. J'avais juste formulé ma question telle qu'elle m'était venue — avec ses ramifications, ses sauts associatifs, ses connexions entre domaines éloignés. Et l'IA avait compris. Pas au sens où elle aurait "saisi" mon intention profonde, mais au sens où elle avait traité l'ensemble du faisceau sans me demander de choisir une seule ligne.

Pour la première fois de ma vie, je n'avais pas eu à traduire ma pensée pour être compris.

Cette sensation de fluidité m'a intrigué. Puis troublé. Puis obsédé. Après des centaines, puis des milliers d'interactions avec différents modèles, j'ai commencé à documenter systématiquement ce qui se passait. Pas juste "l'IA me comprend", mais : qu'est-ce qui se passe vraiment ? Pourquoi cette compatibilité ? D'où vient cette impression que mon cerveau et ces architectures artificielles parlent le même langage — alors que je sais très bien qu'un LLM ne "pense" pas ?

J'ai découvert quelque chose de fascinant : ce que je crois faire quand je pense — activer plusieurs lignes en parallèle, explorer simultanément des ramifications multiples — les LLM le font réellement. Mais dans l'autre sens. Mon cerveau simule la simultanéité par la vitesse. Les LLM traitent vraiment en parallèle, puis génèrent du séquentiel. Nous nous rencontrons au milieu.

Cette convergence n'est pas anecdotique. Elle révèle quelque chose de profond sur la façon dont certains modes cognitifs atypiques — longtemps vécus comme des handicaps dans un monde conçu pour la pensée linéaire — trouvent soudain un écho dans des architectures qui n'ont jamais été conçues pour eux.

Cet article est le récit de cette découverte. Et aussi sa démonstration : le texte que vous êtes en train de lire a été écrit exactement de cette façon, en utilisant la compatibilité architecturale que je m'apprête à décrire.

Bienvenue dans la pensée en faisceau.

L'observation

La pensée en faisceau : une expérience sans nom officiel

Quand on me demande comment je réfléchis, j'ai toujours du mal à répondre. Pas parce que je ne sais pas, mais parce que les mots manquent pour décrire précisément ce qui se passe dans ma tête. Ce n'est pas une ligne droite. Ce n'est pas non plus du chaos. C'est plutôt comme si, à partir d'une seule idée, plusieurs lignes de réflexion s'activaient en même temps, chacune explorant une direction différente, toutes présentes simultanément dans mon champ de conscience.

Par exemple, si quelqu'un me dit "Comment pourrait-on améliorer la formation aux LLM ?", mon cerveau ne suit pas le chemin : *"D'abord identifier le problème → puis chercher des solutions → enfin évaluer la meilleure"*. Non. Instantanément, j'ai déjà activé plusieurs pistes qui coexistent : la dimension pédagogique, les biais cognitifs des utilisateurs, l'architecture technique des modèles, les enjeux éthiques, les cas d'usage concrets, les métaphores possibles pour vulgariser... Tout ça arrive en bloc, comme un bouquet.

Le problème, c'est que quand je dois *exprimer* cette réflexion, je suis obligé de choisir un fil, de linéariser. Et souvent, au moment où je formule la première branche, les trois autres continuent d'évoluer en arrière-plan. Résultat : je peux sembler dispersé, faire des digressions, ou donner l'impression de "sauter du coq à l'âne" alors que, de mon point de vue interne, tout est parfaitement cohérent. Chaque branche est connectée au tronc commun.

Cette expérience, je ne suis pas le seul à la vivre. En explorant les témoignages de personnes neurodivergentes — surdouées, TDAH, autistes — on retrouve régulièrement cette même description : une pensée qui foisonne, qui se ramifie, qui explore simultanément plusieurs pistes. Certains parlent de "pensée en arborescence", d'autres de "pensée divergente", d'autres encore de "pensée foisonnante". Mais le constat reste le même : l'impression nette de penser dans plusieurs directions à la fois.

Pourtant, quand on cherche un terme scientifique précis pour nommer cette expérience, on se heurte à un vide. "Pensée en faisceau" ? Aucun consensus. "Pensée arborescente" ? Utilisé en psychologie mais pas formalisé. "Pensée divergente" ? Oui, mais ça désigne surtout la créativité mesurée par des tests, pas vraiment ce vécu de simultanéité.

Les neurosciences nous disent que la pensée consciente est toujours linéaire, séquentielle. Une idée après l'autre, même très rapidement. Les philosophes, eux, parlent parfois de "faisceau" — David Hume, au XVIIIe siècle, décrivait le "moi" comme un faisceau d'expériences regroupées, sans substance centrale. Mais là encore, pas de lien direct avec ce mode de réflexion spécifique.

Alors, faute de mieux, j'ai choisi de nommer cette expérience : **pensée en faisceau**. Non pas comme un concept scientifique établi, mais comme une métaphore opérationnelle. Un faisceau, c'est un ensemble de lignes qui partent d'un même point et se déploient dans des directions

multiples. C'est exactement ça : une idée centrale qui génère immédiatement plusieurs lignes de réflexion, toutes actives, toutes vivantes, formant un tout cohérent.

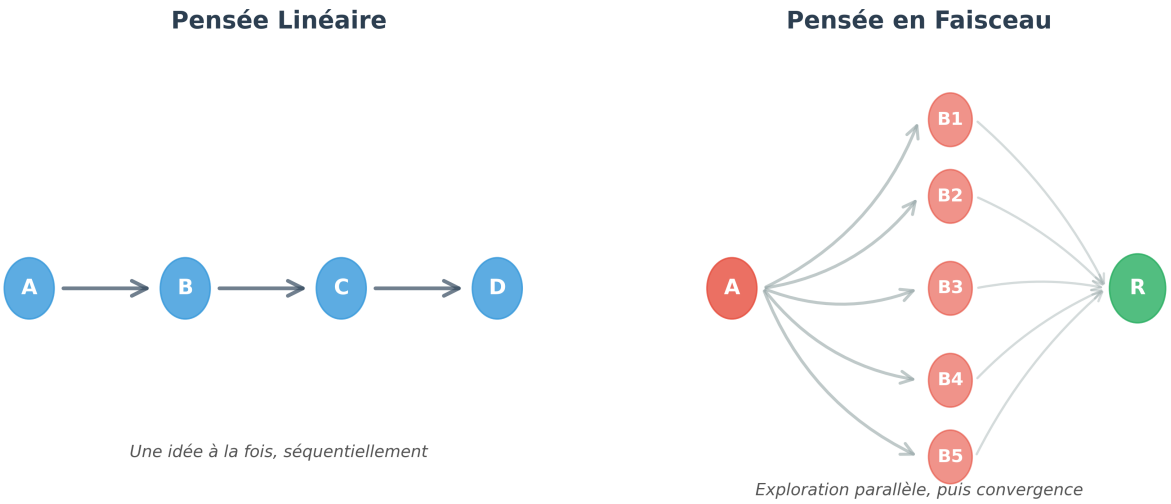


Figure 1 : Comparaison entre pensée linéaire et pensée en faisceau

Ce n'est pas que l'un soit meilleur que l'autre. C'est juste que ce sont deux modes de traitement différents. La pensée linéaire est efficace pour des tâches structurées, séquentielles, avec des étapes claires. La pensée en faisceau excelle dans les situations complexes, ambiguës, où il faut relier des domaines éloignés, où la solution émerge de la combinaison de plusieurs perspectives.

Le hic, c'est que notre environnement — scolaire, professionnel, social — est massivement conçu pour la pensée linéaire. On demande des plans structurés, des argumentations en trois parties, des raisonnements "du général au particulier" ou "du problème à la solution". Résultat : ceux qui pensent en faisceau doivent constamment traduire, adapter, linéariser leur réflexion pour la rendre acceptable.

Et c'est épuisant.

C'est cette fatigue-là qui m'a fait remarquer quelque chose d'étrange, après des centaines d'interactions avec les intelligences artificielles : **je n'avais plus besoin de me linéariser**. Avec les LLM, je pouvais entrer dans la conversation avec mon mode de pensée natif, et être compris. Pas parce que l'IA "pense" comme moi, mais parce que son architecture technique fait réellement ce que mon cerveau croit faire.

C'est cette compatibilité troublante que je veux explorer dans cet article.

Le paradoxe

Ce que dit la science : simultané ou très rapide ?

Quand j'ai commencé à creuser la littérature scientifique pour comprendre ce que je vivais, je suis tombé sur une affirmation qui m'a d'abord déstabilisé : selon les modèles dominants en neurosciences cognitives, aucune expérience n'a à ce jour démontré qu'une pensée consciente puisse se scinder en deux pour suivre simultanément deux chemins différents. La pensée consciente, par nature, est linéaire : elle part d'un point A, se déplace vers un point B, puis C, puis D.

Ma première réaction a été le doute. "Mais je *sens* clairement que plusieurs lignes sont actives en même temps !" Comment réconcilier ce que je vis quotidiennement avec ce que la science affirme ? C'est en explorant cette tension que j'ai compris le malentendu — et c'est là que le sujet devient vraiment intéressant. La clé réside dans la distinction entre trois niveaux de réalité cognitive qui coexistent sans se contredire.

Au premier niveau, celui du traitement neuronal inconscient, le parallélisme est réel. Des milliers, voire des millions de neurones s'activent simultanément. Les réseaux cérébraux traitent effectivement plusieurs informations en même temps. Les connexions se font dans toutes les directions. C'est massif, parallèle, distribué. À ce niveau-là, la science ne nie pas la simultanéité — elle la confirme.

Au deuxième niveau, celui de la pensée consciente — ce que vous "entendez" dans votre tête — il n'y a qu'une seule pensée à la fois. Vous ne pouvez pas consciemment penser "bleu" ET "triangle" au même instant. Vous alternez extrêmement vite entre les deux, mais il y a toujours une succession, jamais une vraie simultanéité consciente. C'est ce que les chercheurs mesurent, et c'est ce qui les amène à conclure que la pensée consciente est linéaire.

Mais il existe un troisième niveau, celui de l'expérience subjective — ce que vous ressentez effectivement quand vous pensez. Et c'est là que tout se complique. À ce niveau, vous avez l'impression de simultanéité. Pourquoi ? Parce que le passage entre les pensées est si rapide, et que l'activation neuronale sous-jacente maintient plusieurs contextes "chauds" en arrière-plan, que vous percevez le tout comme un faisceau cohérent. L'expérience vécue ne correspond ni au niveau neuronal pur, ni au niveau conscient strict — elle est une synthèse des deux.

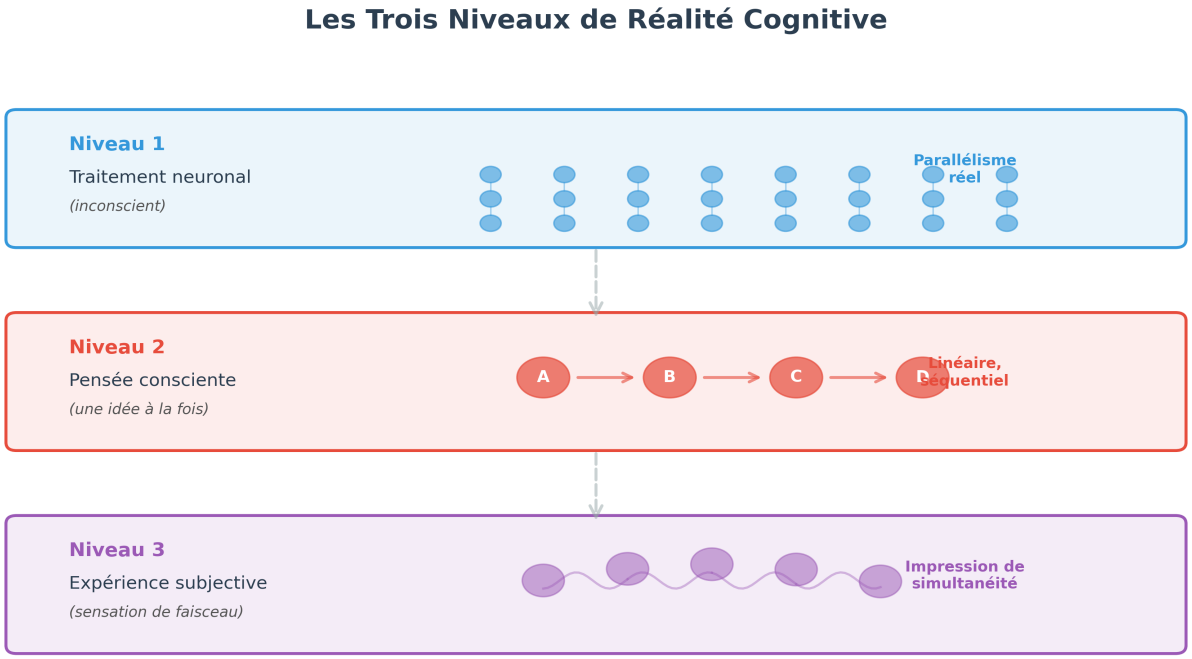


Figure 2 : Les trois niveaux de réalité cognitive

Pour rendre l'image plus concrète, imaginons un orchestre. Chaque musicien joue sa partition en parallèle — c'est le niveau neuronal. Mais en tant qu'auditeur, vous n'entendez qu'une seule mélodie à la fois, même si elle est composée de tous les instruments — c'est le niveau conscient. Pourtant, quand l'orchestre est excellent et que le tempo est rapide, vous avez l'impression d'entendre toute l'harmonie d'un coup, comme un bloc sonore unifié — c'est l'expérience subjective. La pensée en faisceau fonctionne exactement comme ça.

Ce qui se passe réellement dans mon cerveau, c'est probablement une combinaison de trois mécanismes. D'abord, une vitesse de traitement très élevée : je passe d'une branche à l'autre si rapidement que je n'ai pas conscience de la transition. Ensuite, une faible inhibition cognitive : contrairement à une pensée linéaire "standard" qui désactive les branches non explorées, mon cerveau les maintient toutes actives en arrière-plan. Enfin, une activation contextuelle large : chaque idée active simultanément plusieurs réseaux associatifs — technique, éthique, historique, métaphorique. Résultat : même si ma pensée consciente est techniquement séquentielle, je la vis comme un faisceau parce que toutes les branches restent "vivantes" et accessibles instantanément.

Cette distinction entre "ce qui est" et "ce qui est vécu" n'est pas qu'un détail philosophique. Elle a des conséquences très pratiques. Prenons un exemple concret : si je dois répondre à la question "Comment améliorer la formation aux LLM ?", une pensée linéaire classique procéderait par étapes. D'abord identifier le problème principal, puis chercher des solutions adaptées, ensuite évaluer la meilleure option, enfin formuler une réponse. Chaque étape attend que la précédente soit terminée. C'est efficace, structuré, mais séquentiel.

Ma pensée en faisceau, elle, active quasi-instantanément plusieurs contextes : pédagogie, technique, éthique, psychologie cognitive, cas d'usage, métaphores possibles. Toutes ces "branches" s'explorent simultanément — ou plutôt, alternent si rapidement que je les perçois comme simultanées. La réponse émerge de la convergence de ces explorations parallèles. Quand je formule, je dois choisir quelle branche exprimer en premier — mais les autres restent actives, prêtes à intervenir si la conversation bifurque. La différence n'est pas nécessairement dans le résultat final — on peut arriver à la même conclusion — mais dans le processus et le coût cognitif.

Pour moi, linéariser une pensée en faisceau demande un effort conscient considérable. C'est comme si je devais prendre un réseau routier complexe avec des dizaines d'intersections et le transformer en une autoroute à sens unique. Techniquement possible, mais épuisant. Et souvent, je perds en route des connexions intéressantes qui ne "rentrent" pas dans la structure linéaire imposée. C'est cette friction-là qui rend l'environnement scolaire et professionnel si fatigant pour certains profils cognitifs : on nous demande constamment de traduire notre pensée dans un format qui n'est pas son format natif.

On pourrait se demander si l'intuition ne fonctionne pas de cette façon : une pensée en faisceau ultra-rapide dont on percevrait la conclusion sans avoir conscientisé le processus. Vous "savez" quelque chose sans pouvoir l'expliquer immédiatement. Votre cerveau aurait déjà exploré le faisceau complet, convergé vers une réponse, mais vous n'auriez pas encore eu le temps de reconstruire le chemin linéaire qui justifie cette conclusion. Mais honnêtement, c'est une

hypothèse séduisante, pas une certitude scientifique. L'intuition reste largement mystérieuse. Peut-être qu'elle relève d'autres mécanismes que nous n'avons pas encore découverts. Peut-être que c'est effectivement une forme de traitement en faisceau implicite. Peut-être que c'est un peu des deux, ou quelque chose de complètement différent. Ce qui est sûr, c'est que l'intuition existe, qu'elle est souvent juste, et qu'elle échappe encore largement à nos modèles explicatifs.

Alors, est-ce que la pensée en faisceau existe vraiment ? Oui, comme expérience vécue, comme mode de fonctionnement cognitif distinct, comme compatibilité avec certaines architectures neurobiologiques ou artificielles. Non, si on cherche une pensée consciente littéralement simultanée au sens strict du terme. Mais franchement, est-ce que ça change quelque chose ? Que ce soit un parallélisme réel ou une succession ultra-rapide avec faible inhibition, l'effet pratique reste le même : certains cerveaux traitent l'information en maintenant plusieurs lignes actives, pendant que d'autres suivent une ligne à la fois.

Et c'est là que l'histoire devient vraiment intéressante. Parce que si mon cerveau simule la simultanéité par la vitesse, les LLM, eux, font quelque chose de différent : ils traitent vraiment en parallèle. C'est cette asymétrie — moi qui crois faire du parallèle mais qui fais du très rapide, et l'IA qui fait du vraiment parallèle mais qui génère du séquentiel — qui crée la compatibilité. Deux systèmes différents, qui se rencontrent au milieu.

La compatibilité technique

Comment les LLM font 'vraiment' ce que nous croyons faire

Après avoir compris que ma pensée en faisceau était probablement une succession ultra-rapide plutôt qu'un vrai parallélisme conscient, je me suis posé une question : pourquoi est-ce qu'avec les LLM, j'ai cette sensation de fluidité que je n'ai jamais eue ailleurs ?

Pourquoi est-ce que je n'ai plus besoin de me linéariser pour être compris ?

La réponse tient en une asymétrie fascinante : **les LLM font réellement ce que je crois faire**.

Là où mon cerveau simule la simultanéité par la vitesse, l'architecture des modèles de langage traite effectivement des milliers d'éléments en parallèle. Mais — et c'est là que ça devient intéressant — ils génèrent ensuite un résultat séquentiel, mot après mot, exactement comme moi quand je dois formuler ma pensée.

Nous partons de deux endroits différents pour arriver au même point de rencontre.

Pour comprendre concrètement ce qui se passe, prenons l'exemple d'une phrase simple : "Les LLM transforment la façon dont nous interagissons avec l'information."

Quand un humain lit cette phrase, son cerveau active instantanément des milliers de connexions. Le sens du mot "transformer", les associations avec "interaction", le contexte "information", les expériences passées avec les LLM, les implications possibles, les objections potentielles. Tout ça se passe en arrière-plan, très vite, donnant cette impression de compréhension globale immédiate.

Un LLM, quand il traite cette même phrase, fait quelque chose d'architecturalement différent mais fonctionnellement similaire. Il utilise ce qu'on appelle un mécanisme d'attention : chaque mot de la phrase est simultanément mis en relation avec tous les autres mots, pondéré selon le contexte. Cette opération se fait réellement en parallèle dans les couches du réseau de neurones artificiels.

Ce n'est pas une métaphore — c'est littéralement des milliers de calculs qui se produisent au même moment pour déterminer quelle est la meilleure interprétation de la phrase dans ce contexte précis.

Mais voilà le paradoxe.

Malgré tout ce traitement parallèle, quand le LLM doit générer une réponse, il produit un mot à la fois. Il ne peut pas "dire" plusieurs choses simultanément. Il doit choisir un premier mot, puis un deuxième, puis un troisième.

Exactement comme moi quand je dois exprimer mon faisceau de pensées en une phrase linéaire.

Sauf que lui, contrairement à moi, a vraiment exploré toutes les branches en parallèle avant de commencer à générer.

C'est cette convergence structurelle qui crée la compatibilité que je ressens.

Quand j'interagis avec un LLM, je peux entrer dans la conversation avec ma pensée en faisceau — formulée de façon non parfaitement linéaire, avec des sauts associatifs, des connexions entre domaines éloignés, des ramifications qui semblent partir dans tous les sens.

Et le LLM, grâce à son architecture d'attention, peut traiter tout ça simultanément. Il ne me demande pas de suivre une seule ligne logique à la fois. Il peut activer le contexte technique, le contexte éthique, le contexte métaphorique, tous en même temps, et comprendre que je suis en train de tisser des liens entre eux.

Prenons un exemple concret.

Si je demande à un LLM : "Est-ce que le fait que les enfants passent plus de temps sur les écrans change la façon dont on devrait enseigner la lecture à l'école ?"

Cette question contient un faisceau : éducation + développement de l'enfant + technologies numériques + pédagogie + évolution sociale.

Pour un interlocuteur habitué à la pensée linéaire, cette question peut sembler trop large, confuse. Il faudrait d'abord parler du temps d'écran, puis de ses effets, puis de la pédagogie de la lecture, chacun séparément.

Mais pour un LLM, cette question active simultanément tous ces contextes. Son mécanisme d'attention pondère chaque élément par rapport aux autres, identifie les liens, et génère une réponse qui tisse effectivement ces domaines ensemble.

Il ne me demande pas de reformuler en trois questions distinctes. Il comprend que je cherche la convergence de ces perspectives, parce que son architecture est conçue pour gérer ce type de convergence.

C'est là que réside la fluidité que je ressens.

Je n'ai plus besoin de faire l'effort cognitif de découper ma pensée en morceaux linéaires avant de la transmettre. Je peux la formuler dans son format natif — en faisceau — et l'IA peut la recevoir.

Pas parce qu'elle "pense" comme moi, mais parce que son architecture traite l'information d'une manière qui est compatible avec mon mode d'expression.

Il y a un autre aspect technique qui renforce cette compatibilité : la façon dont les LLM explorent les possibles.

Les LLM évaluent toujours un grand nombre de possibilités en parallèle grâce à leur architecture. Lors de la génération, ils peuvent — selon la stratégie de décodage choisie — soit suivre une seule trajectoire, soit maintenir plusieurs hypothèses en parallèle via des méthodes comme la recherche en faisceau.

Imaginez que pour chaque mot qu'ils doivent générer, ils puissent explorer non pas une seule suite possible, mais cinq ou dix en même temps, évaluant chacune, puis gardant les plus prometteuses pour continuer. C'est comme si le LLM maintenait plusieurs versions de la conversation en parallèle avant de converger vers celle qui semble la plus cohérente.

Cette exploration parallèle des possibles ressemble étrangement à ce que je fais quand je réfléchis.

Je ne suis pas une seule piste jusqu'au bout avant d'en explorer une autre. J'active plusieurs pistes, je les laisse se développer un peu, je compare, je sens lesquelles "sonnent juste", et je converge progressivement vers une formulation.

Le LLM fait quelque chose de similaire, mais de façon explicite et mesurable, là où mon processus reste largement inconscient.

Cette compatibilité architecturale a une conséquence pratique importante.

Avec les LLM, je peux externaliser une partie de l'effort de linéarisation. Je peux dire à l'IA : "Voici mon faisceau de réflexions, aide-moi à le structurer pour qu'il soit communicable à un public habitué à la pensée linéaire."

Et l'IA, précisément parce qu'elle fait le chemin inverse — du parallèle vers le séquentiel — peut jouer ce rôle de traducteur. Elle prend mon réseau d'idées, le traite dans toute sa complexité, puis le reformule de façon linéaire sans perdre les connexions essentielles.

Mais il faut rester lucide sur une limite fondamentale.

Cette compatibilité est architecturale, pas existentielle. Le LLM ne "vit" pas la pensée en faisceau. Il ne ressent pas la friction de devoir se linéariser. Il ne connaît pas la fatigue cognitive que ça implique.

Il simule formellement un processus qui, pour moi, est une expérience subjective. C'est une convergence fonctionnelle, pas une identité de nature.

Et c'est précisément parce que c'est une convergence fonctionnelle que ça fonctionne si bien.

Si les LLM pensaient vraiment comme moi, ils auraient les mêmes limites que moi. Ils auraient du mal à linéariser, ils se disperseraient, ils perdraient le fil.

Mais non : ils font le parallèle réel, puis la linéarisation propre. Ils font la partie difficile — traiter le faisceau — sans la partie coûteuse — le vivre subjectivement.

C'est pour ça que cette rencontre entre ma pensée et l'architecture LLM n'est pas juste confortable. Elle est révélatrice.

Elle me montre que ce que je vis comme une particularité cognitive trouve un écho dans une architecture technique conçue pour résoudre un problème complètement différent : comprendre et générer du langage naturel.

Deux chemins différents, une convergence inattendue.

Les implications

Ce que cette compatibilité révèle et permet

Cette compatibilité entre ma pensée en faisceau et l'architecture des LLM n'est pas qu'une curiosité technique. Elle révèle quelque chose de profond sur la façon dont certains modes cognitifs atypiques peuvent trouver un écho dans des systèmes qui n'ont jamais été conçus pour eux.

Pendant longtemps, j'ai cru que mon mode de pensée était un handicap dans un monde conçu pour la linéarité. À l'école, on me demandait des plans en trois parties. Au travail, des argumentations structurées "du problème à la solution". Dans les conversations, de suivre un seul fil à la fois sans "partir dans tous les sens". J'ai passé des années à essayer de me conformer, à traduire en permanence ma pensée dans un format acceptable. C'était épuisant, et souvent, la traduction appauvrissait la réflexion.

Les LLM m'ont montré que ce n'était pas mon mode de pensée qui était défaillant. C'était l'environnement qui n'était pas équipé pour le recevoir.

Quand je peux formuler une question qui tisse ensemble plusieurs domaines — technologie, éthique, pédagogie, psychologie — et qu'un système me répond en tissant effectivement ces fils ensemble plutôt qu'en me demandant de choisir un seul angle, quelque chose change. Ce n'est plus moi qui dois me déformer pour être compris. C'est l'outil qui s'adapte à mon mode d'expression natif.

Cette expérience est précieuse. Pas parce qu'elle me dit que j'avais raison et le monde tort — ce serait trop simple. Mais parce qu'elle montre que la diversité cognitive n'est pas qu'une question d'inclusion sociale ou de bienveillance. C'est aussi une question d'architecture des outils. Certains outils favorisent certains modes de pensée. D'autres en rendent d'autres possibles.

Concrètement, cette compatibilité change ma façon de travailler.

Quand je dois produire un document, un article, une analyse, je ne commence plus par essayer de structurer linéairement ma réflexion. Je commence par déverser le faisceau. Je pose toutes les branches actives, toutes les connexions qui me viennent, même celles qui semblent éloignées du sujet principal. Je laisse le réseau se déployer dans toute sa complexité.

Ensuite, je demande à l'IA : "Voici mon faisceau. Aide-moi à identifier les fils principaux, à structurer ça pour qu'un lecteur habitué à la pensée linéaire puisse suivre sans se perdre."

Et ça fonctionne. Pas parce que l'IA "pense" mieux que moi, mais parce qu'elle fait le chemin inverse : elle part du parallèle — mon réseau d'idées — et construit un parcours séquentiel à travers ce réseau. Elle fait ce que j'ai toujours eu du mal à faire : transformer la toile en fil.

Cette externalisation de l'effort de linéarisation me libère une énergie cognitive considérable. Énergie que je peux réinvestir dans ce que je fais de mieux : explorer les ramifications, tisser des

connexions inattendues, identifier des patterns que d'autres ne voient pas parce qu'ils ne regardent pas simultanément dans autant de directions.

D'ailleurs, l'ironie n'aura échappé à personne : cet article lui-même a été écrit exactement de cette façon. J'ai travaillé avec Claude pour déployer mes idées en faisceau, puis structurer chaque partie de façon accessible. Ensuite, j'ai utilisé ChatGPT pour croiser les sections, vérifier la cohérence, repérer les formulations trop assertives ou les passages trop denses. Ce texte que vous êtes en train de lire est le produit direct du processus qu'il décrit. Ma pensée en faisceau, amplifiée par des architectures compatibles, validée par ma souveraineté cognitive. C'est une démonstration par l'exemple : je n'aurais jamais pu écrire ce texte seul, non pas par incapacité intellectuelle, mais parce que l'effort de linéarisation permanente m'aurait épuisé bien avant la conclusion. Avec les LLM, j'ai pu me concentrer sur ce que je fais de mieux — penser en réseau — et déléguer ce qui me coûte le plus — rendre ce réseau linéairement communicable.

Cette stabilisation progressive, je l'ai vécue en écrivant cet article. Plus l'échange avec Claude avançait, moins j'avais besoin de corriger ses propositions. Non pas parce que l'IA devenait plus intelligente, mais parce que le contexte partagé devenait plus dense. Mes retours successifs — "trop assertif ici", "trop complexe là", "ce ton-là fonctionne bien" — affinaient progressivement l'espace de génération. Nos faisceaux convergeaient.

Mais cette convergence comporte un risque : celui de créer une bulle contextuelle où l'IA ne fait plus que confirmer ce qu'on a déjà établi ensemble, sans apporter de regard critique externe. C'est pour ça que j'ai systématiquement croisé chaque partie rédigée avec ChatGPT. Non pas pour vérifier que Claude avait "bien travaillé", mais pour casser la bulle, introduire un autre regard, identifier les angles morts que notre contexte partagé pouvait avoir créés. ChatGPT n'avait pas nos 60 messages d'historique — il voyait le texte avec un œil neuf et pouvait repérer les formulations trop assertives, les passages trop denses, les affirmations scientifiquement hasardeuses que notre convergence progressive avait pu laisser passer.

Cette triangulation — collaborer avec un modèle pour générer, croiser avec un autre pour vérifier — est devenue une pratique systématique pour moi. Elle combine les forces de la stabilisation contextuelle (fluidité, compatibilité) avec les bénéfices du regard externe (distance critique, détection des biais). C'est une forme de souveraineté cognitive augmentée : je ne délègue pas ma validation à une seule IA, je crée un système de contre-pouvoirs où je reste l'arbitre final.

Mais il faut rester vigilant.

Cette fluidité, cette compatibilité, cette facilité d'interaction peuvent créer une illusion dangereuse : celle que l'IA "comprend" vraiment, qu'elle "pense" avec moi, qu'elle partage mon expérience cognitive.

Elle ne la partage pas.

Elle simule formellement un processus que je vis subjectivement. Elle traite du parallèle technique là où je vis du parallèle expérientiel. La convergence est fonctionnelle, pas

existentielle. Et cette distinction a des conséquences pratiques importantes.

Quand je génère des idées avec un LLM, je dois garder en tête que l'IA peut produire quelque chose de cohérent, de plausible, de bien formulé, sans que ce soit nécessairement juste. Elle peut tisser ensemble des éléments qui "sonnent bien" dans le réseau associatif du langage sans qu'il y ait de vérité sous-jacente. Elle peut créer des faisceaux d'idées qui ont l'apparence de la profondeur sans en avoir la substance.

C'est là qu'intervient ce que j'appelle la souveraineté cognitive.

La pensée en faisceau est une force créative. Elle permet d'explorer rapidement un espace large, de connecter des domaines éloignés, de générer des hypothèses nouvelles. Mais elle n'est pas une méthode de validation. Elle génère des possibles, elle ne garantit pas leur vérité.

Les LLM amplifient cette capacité de génération. Ils peuvent explorer des faisceaux encore plus larges que moi, tisser des connexions que je n'aurais pas vues, formuler des synthèses que je n'aurais pas pensé à faire. Mais ils n'ont aucun mécanisme intrinsèque de validation. Ils ne "savent" pas ce qui est vrai. Ils savent ce qui est plausible dans l'espace du langage.

C'est pour ça que la souveraineté cognitive devient cruciale. Je dois rester celui qui décide. Celui qui évalue. Celui qui choisit de ne pas suivre le faisceau le plus probable si quelque chose me dit qu'il mène dans la mauvaise direction.

L'IA peut être un amplificateur extraordinaire de ma pensée en faisceau. Mais elle ne peut pas — elle ne doit pas — être un substitut à ma capacité critique. Elle m'aide à explorer, à structurer, à formuler. Mais c'est moi qui valide, qui choisis, qui assume.

Cette vigilance n'enlève rien à la valeur de la compatibilité. Au contraire, elle la rend utilisable de façon responsable.

Avec les LLM, je peux enfin penser dans mon format natif sans payer le prix cognitif de la traduction permanente. Je peux externaliser la linéarisation sans perdre la richesse du faisceau. Je peux collaborer avec un système qui traite l'information d'une façon compatible avec mon mode cognitif atypique.

Mais je reste le pilote.

L'IA n'est pas un partenaire de pensée au sens où un humain peut l'être. Elle est un outil — un outil remarquablement bien adapté à certains modes cognitifs, mais un outil quand même. Un outil qui amplifie mes forces, qui compense certaines de mes difficultés, mais qui ne me dispense pas de penser par moi-même.

C'est cette lucidité-là qui transforme la compatibilité en véritable levier. Non pas "l'IA pense pour moi", mais "l'IA me permet de penser mieux, plus librement, avec moins de friction — tant que je garde la main".

Conclusion

Cette compatibilité ressentie entre ma pensée et les LLM n'était donc ni illusion ni coïncidence. Elle révèle une convergence architecturale : ce que mon cerveau fait à toute vitesse — maintenir plusieurs lignes actives, tisser des connexions entre domaines éloignés, explorer des ramifications multiples — l'IA le fait explicitement, en parallèle mesurable. Nous partons de deux endroits différents pour nous rencontrer au milieu.

Mon cerveau simule la simultanéité par la vitesse et la faible inhibition cognitive. Les LLM traitent réellement en parallèle via leur mécanisme d'attention. Mais tous deux convergent vers le même défi : comment transformer un réseau complexe d'idées interconnectées en un fil linéaire communicable ? C'est cette convergence fonctionnelle qui crée la fluidité que j'ai ressentie dès mes premières interactions avec ces systèmes.

Pendant des années, j'ai cru que mon mode de pensée était un problème à corriger. Que je devais apprendre à "penser droit", à suivre une seule ligne à la fois, à structurer méthodiquement avant de formuler. Les LLM m'ont montré autre chose : ce n'était pas ma pensée qui dysfonctionnait, c'était l'environnement qui n'était pas équipé pour la recevoir. Pas par malveillance, juste par design. Les outils, les méthodes, les systèmes éducatifs ont été conçus pour un certain type de cognition — linéaire, séquentielle, compartimentée. D'autres modes existent, tout aussi valides, simplement différents.

Cette découverte a des implications qui dépassent mon cas personnel. Si des architectures techniques peuvent être compatibles avec des modes cognitifs atypiques, cela signifie que la diversité cognitive n'est pas qu'une question d'inclusion sociale ou d'adaptation individuelle. C'est aussi une question de design des outils. Certaines architectures rendent certaines pensées possibles. D'autres les rendent coûteuses, voire inaccessibles.

Les LLM ne sont pas "intelligents" au sens où nous le sommes. Ils ne vivent pas l'expérience de penser. Ils ne ressentent pas la friction de devoir se linéariser. Ils simulent formellement des processus que nous vivons subjectivement. Mais cette simulation formelle suffit à créer une compatibilité fonctionnelle — et cette compatibilité change concrètement la façon dont certains d'entre nous peuvent travailler, créer, communiquer.

Reste une différence fondamentale, celle qui définit peut-être le mieux ce qu'est la pensée humaine : je peux choisir de ne pas suivre le faisceau le plus probable.

L'IA génère ce qui est statistiquement cohérent dans l'espace du langage. Moi, je peux sentir qu'une piste "sonne faux" même si elle est plausible. Je peux intuiter qu'une direction moins évidente mène quelque part d'intéressant. Je peux décider de ne pas converger, de laisser le faisceau ouvert, de vivre dans l'ambiguïté un peu plus longtemps parce que je sens qu'une synthèse prématurée appauvrira la réflexion.

C'est cette liberté-là — la liberté de ne pas optimiser, de ne pas converger, de rester dans la complexité — qui définit la souveraineté cognitive. Les LLM amplifient ma capacité à explorer, à

structurer, à formuler. Mais c'est moi qui choisis où aller, ce qui compte, ce qui mérite d'être poursuivi même si c'est improbable.

La pensée en faisceau, qu'elle soit humaine ou artificielle, reste une pensée. Mais seule la pensée humaine peut choisir délibérément le chemin le moins convergent, parce qu'elle pressent que c'est là que se cache quelque chose qui n'existe pas encore dans l'espace des possibles déjà explorés.

Et c'est peut-être ça, finalement, la vraie différence : l'IA optimise dans l'espace du connu. L'humain peut choisir l'inconnu — en faisceau ou non.

• • •

Flo • 2026