DOSSIER III · TEXTE LONG

La métacognition : penser sa propre pensée

Mesure, cerveau, apprentissage, santé mentale, décision collective, expertise, intelligence artificielle et débat public : ce que coûte une confiance mal calibrée.

Franck Herlange · environ 7 000 mots · 32 min de lecture · mis à jour en novembre 2026

Synthèse

La métacognition — la capacité à surveiller et à réguler sa propre pensée — est une fonction distincte de la compétence elle-même. Elle a une base cérébrale identifiable, mûrit jusqu'à la fin de l'adolescence, varie fortement d'un adulte à l'autre et s'enseigne. Ses écarts pèsent sur l'apprentissage, la santé mentale, la décision collective et, plus indirectement, l'insertion sociale.

Huit constats structurent le dossier :

- **Elle se mesure à part de la performance.** La justesse de la confiance — savoir quand on a raison ou tort — se quantifie indépendamment du niveau de réussite.
- **Elle a un siège cérébral.** La capacité d'introspection est corrélée au volume du cortex préfrontal antérieur, région très développée chez l'être humain.
- **Elle apparaît tôt et mûrit tard.** Dès 20 mois, un enfant demande de l'aide quand il ne sait pas ; la capacité progresse ensuite jusqu'à 17 ans.

- **Elle trompe l'apprenant.** Dans des expériences contrôlées, plus de 80 % des participants jugent la méthode qui donne l'impression d'apprendre au moins aussi efficace que celle qui fait réellement apprendre ; enseigner la métacognition apporte en moyenne 8 mois de progrès scolaire en un an.
- **Ses dérèglements traversent la santé mentale.** Rumination, sous-confiance lucide, surconfiance aveugle et réflexivité faible en sont des formes ; des thérapies les ciblent.
- **Elle conditionne la coopération.** Quand l'un des deux partenaires est nettement moins performant, la décision commune devient pire que celle du meilleur seul, et la surconfiance procure du statut : un mécanisme plausible d'exclusion, encore peu documenté sur le terrain de l'emploi.
- **L'expertise n'en protège pas.** Des médecins dont la justesse tombe de 55 % à 6 % ne perdent presque rien de leur confiance ; les fourchettes à 80 % de directeurs financiers ne contiennent le résultat que 36 % du temps. Une heure de formation améliore pourtant de 6 à 11 % la justesse de prévisionnistes.
- **Elle devient une compétence clé à l'ère de l'IA et du débat public.** Des explications plus longues font davantage croire aux réponses d'une IA sans les rendre plus justes ; à l'inverse, devoir expliquer le mécanisme d'une politique publique fait baisser la compréhension ressentie et modère les positions.

1. Définitions : ce que recouvre la métacognition

La métacognition est la connaissance et le contrôle qu'un individu exerce sur ses propres processus de pensée. Le terme est fixé par John Flavell en 1979, qui la décrit comme « la connaissance et la cognition portant sur les phénomènes cognitifs » (Flavell, 1979).

Flavell distingue deux composantes, toujours structurantes aujourd'hui (résumé APA) :

- **Les connaissances métacognitives** : ce que l'on sait ou croit de soi et des autres comme agents pensants, des tâches et des stratégies.
- **Les expériences métacognitives** : les ressentis conscients, cognitifs ou affectifs, qui surviennent pendant une activité — typiquement, le sentiment que « ça se passe bien ou mal ».

La littérature ultérieure retient un couple opérationnel : **surveillance** (évaluer où l'on en est) et **contrôle** (ajuster sa conduite en conséquence).

Nelson et Narens ont formalisé ce couple en 1990 avec deux niveaux : un niveau « objet » — mémoire, perception, raisonnement — et un niveau « méta » qui en tient un modèle. La surveillance fait remonter l'information de l'objet vers le méta; le contrôle redescend pour modifier le processus : poursuivre une recherche en mémoire, l'abandonner, consacrer plus de temps à une leçon (Nelson & Narens, 1990).

La dimension sociale : de soi aux autres

Le point décisif pour l'insertion est que la métacognition ne porte pas seulement sur soi. Chris Frith montre que le même mécanisme, appliqué à autrui, s'appelle la **mentalisation** (Frith, 2012, Phil. Trans. R. Soc. B). Il distingue deux étages :

Forme	Fonctionnement	Rôle social
Implicite	Automatique, sans conscience	Permet le « mode nous » : intégrer d'emblée les savoirs et intentions d'autrui dans l'action commune
Explicite	Délibérée, verbalisable	Permet de justifier sa conduite et de délibérer à plusieurs

Frith soutient que la métacognition explicite s'est développée parce qu'elle améliore la décision collective. Conséquence pour ce dossier : un fonctionnement métacognitif qui s'écarte de la norme — par défaut, par excès ou par décalage — peut toucher la capacité à coopérer, et par là l'insertion ; les sections 8 à 12 disent jusqu'où cette hypothèse est établie.

2. Mesurer la métacognition

Longtemps, la métacognition a été étudiée par questionnaire : on demandait aux gens s'ils se jugeaient bons à telle ou telle tâche. Ces déclarations mesurent une image de soi, pas la capacité réelle à juger sa pensée. La recherche actuelle procède autrement : elle compare, essai après essai, la confiance qu'une personne déclare et la justesse effective de ses réponses. Une bonne métacognition se reconnaît à ceci : quand on est sûr, on a plus souvent raison (Fleming & Lau, *Frontiers in Human Neuroscience*, 2014).

Deux niveaux de jugement

Le cadre de référence vient de la théorie de la détection du signal, qui distingue deux types de jugement :

- **Le jugement de premier ordre** porte sur le monde : ce signal est-il présent ? cette réponse est-elle A ou B ? Sa qualité mesure la performance.

- **Le jugement de second ordre** porte sur le premier : ma réponse était-elle juste ? Sa qualité mesure la métacognition.

Une personne peut être excellente au premier niveau et médiocre au second, ou l'inverse. C'est cette indépendance de principe qui fait de la métacognition une capacité à part entière, et non un simple reflet de la compétence.

Trois grandeurs à ne pas confondre

Fleming et Lau insistent sur une distinction que le langage courant confond sous le mot « confiance » :

Mesure	Question posée	Exemple d'écart
Biais	Quel est le niveau moyen de confiance ?	Une personne systématiquement trop sûre, ou jamais assez
Sensibilité	La confiance distingue-t-elle les bonnes réponses des erreurs ?	Une confiance identique, qu'on ait raison ou tort
Efficacité	Quelle sensibilité, à performance égale ?	Deux personnes aussi performantes, dont une seule sait quand elle se trompe

Le biais et la sensibilité varient indépendamment. Une personne très prudente peut afficher une confiance basse en moyenne tout en ajustant finement cette confiance à la qualité de chaque réponse : son biais est négatif, sa sensibilité excellente. À l'inverse, une personne toujours assurée peut l'être autant quand elle se trompe que quand elle a raison : biais positif, sensibilité nulle. Lorsque les jugements sont exprimés en probabilités, on peut tracer des courbes de calibration qui rendent visible la surconfiance ou la sous-confiance.

Isoler la métacognition de la performance

Les premiers indices, comme le gamma proposé par Nelson en 1984 ou l'aire sous la courbe de second ordre, ont un défaut : ils dépendent de la performance. Une personne qui réussit mieux paraît mécaniquement mieux calibrée. La mesure dite meta-d', proposée par Maniscalco et Lau, contourne ce problème en exprimant la sensibilité métacognitive dans la même unité que la performance. Le rapport meta-d'/d' donne alors l'efficacité métacognitive : à niveau de réussite donné, quelle part de l'information disponible la personne exploite-t-elle pour juger ses propres réponses ? (Fleming, 2017, présentation de la méthode).

Ce progrès méthodologique conditionne la plupart des résultats récents cités dans ce dossier. Sans lui, impossible d'affirmer qu'un trait psychologique modifie la métacognition plutôt que la performance.

3. Bases biologiques et évolutives

La métacognition n'est pas une abstraction philosophique : elle a un support cérébral identifiable et des précurseurs chez certains animaux.

Un siège dans le cortex préfrontal antérieur

Fleming et al. ont fait passer à des volontaires une tâche de perception visuelle dont la difficulté était ajustée pour que tous réussissent au même niveau (*Science*, 2010). Après chaque réponse, les participants indiquaient leur confiance. Les chercheurs ont ainsi pu isoler la capacité d'introspection de chacun, puis la comparer à l'anatomie de son cerveau mesurée par IRM.

Trois constats en ressortent :

- **Une région dédiée.** La capacité d'introspection est corrélée au volume de matière grise du cortex préfrontal antérieur droit (aire de Brodmann 10). Cette

région a connu un développement évolutif marqué chez l'être humain.

- **Des connexions associées.** La microstructure de la matière blanche reliée à cette zone varie aussi avec la capacité métacognitive : ce n'est pas seulement la région, mais son câblage, qui compte.
- **Une indépendance vis-à-vis de la performance.** Les participants réussissaient la tâche au même niveau ; seule leur capacité à juger leurs réponses différait, et c'est elle que l'anatomie prédisait (UCL, 2010).

Ces résultats rejoignent des observations cliniques plus anciennes, rappelées dans l'article : des lésions préfrontales s'accompagnent d'une introspection dégradée. Ils suggèrent que juger sa propre décision mobilise un substrat distinct de celui qui produit la décision.

L'étude établit une corrélation, pas un déterminisme. Elle ne dit pas si cette structure est innée ou façonnée par l'exercice ; la question reste ouverte.

Des précurseurs chez l'animal

Smith, Shields et Washburn ont adapté des tâches perceptives et de mémoire à des singes rhésus et à un dauphin (*Behavioral and Brain Sciences*, 2003). L'innovation consistait à offrir une réponse « incertain » permettant de décliner un essai difficile. Les animaux l'ont utilisée selon un schéma essentiellement identique à celui des humains placés dans la même situation (Université de Buffalo, 2003).

L'interprétation est discutée. Certains commentateurs estiment que ces conduites montrent une surveillance de l'incertitude sans prouver une représentation consciente de sa propre pensée. Le résultat établit au minimum que la régulation de l'action par un signal d'incertitude est ancienne dans l'évolution, et que la métacognition humaine s'appuie sur ce socle.

4. Développement : du nourrisson à l'adulte

La métacognition apparaît bien plus tôt qu'on ne le pensait, mais elle mûrit lentement, jusqu'à la fin de l'adolescence.

Âge	Ce que montre la recherche	Source
20 mois	Des nourrissons demandent de l'aide à leur parent quand ils ont oublié où un jouet est caché, et améliorent ainsi leurs résultats	Goupil, Romand-Monnier & Kouider, *PNAS*, 2016
7–12 ans	Les jugements de confiance sur ses souvenirs s'affinent	Weil et al., 2013 (revue de littérature)
11–17 ans	La capacité métacognitive progresse alors que la performance reste stable ; elle culmine en fin d'adolescence	Weil et al., *Consciousness and Cognition*, 2013
8–17 ans	Les adolescents s'appuient moins sur les conseils d'autrui, et cette autonomie s'explique par leurs progrès métacognitifs	Moses-Payne et al., 2021
Âge adulte	La capacité plafonne ; l'âge affecte ensuite la performance, mais pas la métacognition	Rouault et al., 2018

Avant le langage : savoir qu'on ne sait pas

On a longtemps cru la métacognition tardive, parce qu'on la mesurait par des comptes rendus verbaux. L'équipe de Sid Kouider à l'ENS a contourné l'obstacle avec un test sans paroles. Quatre-vingts enfants de 20 mois devaient retrouver un jouet caché sous l'une de deux boîtes après un délai variable ; une partie d'entre eux avait appris qu'ils pouvaient se tourner vers leur parent pour obtenir de l'aide (Goupil et al., 2016).

Les enfants disposant de cette option l'ont utilisée de façon sélective, pour éviter les erreurs et décliner les choix difficiles. Leur taux de bonnes réponses a atteint 66 %, contre 56 % dans le groupe sans option. Les auteurs en concluent que des enfants

qui ne parlent pas encore surveillent déjà leur propre incertitude et savent la communiquer. La même équipe a publié la même année des indices neuronaux de sensibilité métacognitive chez le nourrisson préverbal (Goupil & Kouider, *Current Biology*, 2016).

L'adolescence, période de maturation

Weil et al. ont testé 56 personnes âgées de 11 à 41 ans sur une tâche qui sépare performance et métacognition (2013). Entre 11 et 17 ans, la performance reste stable mais la métacognition progresse nettement; elle culmine en fin d'adolescence puis plafonne à l'âge adulte. Cette trajectoire coïncide avec la maturation prolongée du cortex préfrontal antérieur, la région identifiée en section 3.

Cette maturation a des effets concrets. Moses-Payne et al. ont comparé des enfants de 8–9 ans et des adolescents de 12–13 et 16–17 ans : les plus âgés s'appuient moins sur les conseils d'autrui, et ce recul s'explique par leurs progrès métacognitifs (2021). Mieux juger sa propre fiabilité permet de décider seul quand on est fiable, et de solliciter les autres quand on ne l'est pas.

Dès l'origine, la métacognition est donc à la fois une affaire intime et une affaire sociale : savoir qu'on ne sait pas sert d'abord à demander de l'aide.

5. Différences individuelles : être compétent et mal s'évaluer

À âge et performance égaux, les adultes diffèrent nettement par leur métacognition. Deux programmes de recherche, à vingt ans d'écart, éclairent ces différences : le premier montre que l'incompétence et l'aveuglement sur soi vont souvent ensemble; le second, plus fin, montre que la métacognition varie selon des traits psychologiques stables, indépendamment de la performance.

Les moins compétents se surestiment

Kruger et Dunning ont soumis des étudiants à des tests d'humour, de grammaire et de logique, puis leur ont demandé d'estimer leur score et leur rang (*JPSP*, 1999). Les participants du quartile inférieur se plaçaient en moyenne au 62e percentile, alors que leurs résultats les situaient au 12e.

Les auteurs parlent d'un double fardeau : les moins compétents commettent des erreurs, et cette même incompétence les prive de l'outil qui leur permettrait de les voir. Plusieurs analyses rattachent cet écart à un déficit de la capacité à distinguer ses réponses justes de ses erreurs, c'est-à-dire à un déficit métacognitif.

Ce résultat, devenu célèbre sous le nom d'effet Dunning-Kruger, est aussi contesté : une partie de la littérature l'attribue à un artefact statistique (voir section 14). Il reste une illustration utile d'un point central : la confiance qu'on a en soi n'est pas un bon indicateur de sa compétence.

La métacognition se sépare de la performance

L'étude de Rouault et al. apporte une réponse plus rigoureuse, grâce aux méthodes de mesure décrites en section 2 (*Biological Psychiatry*, 2018). Près de 1 000 personnes de la population générale (N = 995), recrutées en ligne, ont effectué une tâche perceptive avec jugements de confiance et rempli des questionnaires sur un large éventail de traits psychologiques, de l'impulsivité à l'anxiété.

Le résultat principal est une double dissociation : les traits psychologiques prédisent des variations de la métacognition mais pas de la performance, tandis que l'âge prédit des variations de la performance mais pas de la métacognition. Autrement dit, ce qui distingue psychologiquement les individus se loge dans leur façon de juger leurs réponses, non dans leur capacité à répondre.

Deux profils opposés en ressortent, que l'analyse par diagnostic séparé n'aurait pas fait apparaître (British Psychological Society, 2018) :

Dimension	Confiance moyenne	Sensibilité métacognitive	Traduction concrète
Anxiété et dépression	Plus basse	Plus élevée	Doute davantage, mais sait mieux quand il se trompe
Comportements compulsifs et pensées intrusives	Plus haute	Plus basse	Se sent plus sûr, mais distingue moins bien ses erreurs

Ce que cela change

Ces résultats défont une équivalence intuitive entre assurance et lucidité. Deux personnes également compétentes peuvent se présenter de façon opposée : l'une paraîtra hésitante alors qu'elle est lucide, l'autre assurée alors qu'elle s'évalue mal. Dans toute situation où l'on juge autrui sur la confiance qu'il affiche — un examen oral, un entretien, une réunion —, cet écart produit des erreurs d'appréciation systématiques. La section 8 montre que les groupes humains y sont particulièrement exposés.

6. Apprendre : illusions et enseignement

L'apprentissage est le terrain où la métacognition a les effets les plus mesurables, dans les deux sens : elle trompe l'apprenant quand elle se fie à de faux signaux, et elle s'enseigne avec un profit parmi les plus élevés de toutes les interventions éducatives.

Le sentiment d'apprendre n'est pas l'apprentissage

Kornell et Bjork ont fait étudier six tableaux de chacun de douze peintres peu connus, soit regroupés par artiste, soit entremêlés avec ceux des autres (*Psychological Science*, 2008). Les participants devaient ensuite attribuer des tableaux inédits au bon peintre. L'étude entremêlée donnait de nettement meilleurs résultats ; les participants jugeaient pourtant l'étude groupée plus efficace, même après avoir constaté le contraire dans leurs propres résultats.

Sur l'ensemble de ces expériences, plus de 80 % des participants jugeaient le regroupement au moins aussi efficace, alors que 15 % seulement réussissaient effectivement mieux ainsi (bilan dressé par Kornell et al., 2010). Une étude de 2010 retrouve le même avantage de l'entremêlement, et la même illusion, chez les adultes âgés : 75 % d'entre eux jugeaient le regroupement supérieur, 4 % seulement l'entremêlement (Kornell et al., *Psychology and Aging*, 2010).

Le mécanisme est général : la fluidité ressentie pendant l'étude est prise pour un signe de maîtrise. Regrouper donne une impression de facilité ; entremêler donne une impression de difficulté, alors que c'est cette difficulté qui fait apprendre.

Les méthodes préférées sont souvent les moins efficaces

Dunlosky et al. ont passé en revue dix techniques d'étude courantes (*Psychological Science in the Public Interest*, 2013) :

Utilité	Techniques
Élevée	S'entraîner par des tests ; étaler l'étude dans le temps
Modérée	S'interroger sur le pourquoi ; s'expliquer à soi-même ; entremêler les exercices
Faible	Résumer ; surligner ; utiliser des moyens mnémotechniques ; se représenter le texte en images ; relire

La plupart des étudiants déclarent relire et surligner, alors que ces techniques n'améliorent pas régulièrement les résultats. Les auteurs recommandent de remplacer la relecture par l'entraînement à se tester. Cet écart entre les méthodes choisies et les méthodes efficaces est une illusion métacognitive à l'échelle de toute une population d'apprenants. Dunlosky y voit en partie un effet de la formation des enseignants : ces techniques sont largement absentes dans les manuels de psychologie de l'éducation qu'ils étudient (Kent State University, 2013).

Progresser rend plus lucide

Kruger et Dunning observent un résultat paradoxal : améliorer les compétences des participants améliore aussi leur capacité à reconnaître leurs limites (Kruger & Dunning, 1999). Dans un domaine donné, compétence et métacognition progressent ensemble. L'aveuglement sur ses lacunes n'est donc pas une fatalité : il recule avec l'apprentissage lui-même.

Enseigner la métacognition : l'intervention la mieux notée

La Education Endowment Foundation, organisme britannique d'évaluation des pratiques éducatives, classe l'enseignement de la métacognition et de l'autorégulation parmi les interventions à fort impact pour un coût très faible, avec un niveau de preuve jugé élevé (EEF, *Teaching and Learning Toolkit*). L'effet moyen est de 8 mois de progrès supplémentaires sur une année scolaire, dans la version actuelle de la synthèse.

L'EEF précise les conditions de cet effet :

- **Un effet difficile à obtenir en pratique**, parce que ces méthodes exigent des élèves qu'ils prennent davantage la responsabilité de leur apprentissage.

- **Une efficacité maximale quand les stratégies sont intégrées** au programme et aux cours d'une discipline, plutôt qu'enseignées à part.
- **Une dépendance à la formation des enseignants** : l'efficacité varie selon leur compréhension de la manière de développer les connaissances métacognitives des élèves, et l'essentiel du coût tient à cette formation.

7. Santé mentale : quand la métacognition se dérègle

Les distorsions de la métacognition sont un trait transversal de la souffrance psychique. Rouault et al. concluent à un lien spécifique et étendu entre métacognition et santé mentale (2018). Ce lien prend des formes opposées selon les troubles, ce qui en fait une cible thérapeutique à part entière.

Trop ou trop peu de réflexivité

Dérèglement	Ce que montre la recherche	Source
Réflexivité emballée	Un style persévérant — inquiétude, rumination, surveillance de la menace — entretenu par des croyances sur ses propres pensées	Wells, *Frontiers in Psychology*, 2019
Confiance basse, lucidité haute	Profil anxieux-dépressif dans la population générale	Rouault et al., 2018
Confiance haute, lucidité basse	Profil compulsif et à pensées intrusives	Rouault et al., 2018
Réflexivité faible	Difficulté à relier pensées et ressentis à leurs causes, étudiée surtout dans la schizophrénie	Lysaker et al., 2020

Le modèle de Wells : quand penser sur sa pensée entretient la souffrance

Le modèle S-REF, proposé par Wells et Matthews en 1994, renverse la perspective habituelle (Wells, 2019). Ce n'est pas d'abord le contenu des pensées négatives qui fait souffrir, mais la manière dont on s'y rapporte. Le modèle décrit un « syndrome

cognitivo-attentionnel » : un style de traitement persévérant fait d'inquiétude, de rumination et de surveillance de la menace.

Ce syndrome est entretenu par deux familles de croyances métacognitives :

- **Des croyances positives**, qui font de l'inquiétude un outil : « m'inquiéter m'aide à me préparer ».
- **Des croyances négatives**, qui la rendent incontrôlable ou dangereuse : « je ne contrôle pas mes pensées ».

Le premier type pousse à ruminer ; le second empêche d'arrêter. La thérapie métacognitive vise ces croyances plutôt que le contenu des pensées.

La schizophrénie : la réflexivité comme pont vers la vie sociale

Le groupe de Paul Lysaker étudie une forme large de métacognition : la capacité à former une représentation intégrée de soi et des autres, en reliant pensées, émotions et événements. Un numéro spécial de *Schizophrenia Research: Cognition* (2020) la relie à l'emploi, à l'empathie, à la motivation et à l'auto-évaluation (Lysaker et al., 2020).

Deux résultats ressortent. La métacognition y apparaît comme un nœud central entre les capacités cognitives, les facteurs sociaux et le fonctionnement réel : elle module l'effet des premiers sur le second. Et elle n'est pas figée : sur 75 personnes suivies 12 mois, elle s'améliore significativement (Kukla & Lysaker, 2020).

Trois thérapies ciblées

Intervention	Principe	Preuves disponibles
Entraînement métacognitif (MCT, Moritz)	8 à 10 modules en groupe : psychoéducation, correction des biais de raisonnement, stratégies	Méta-analyse de 43 études (N = 1 816) : effets sur les idées délirantes (g = 0,69), le fonctionnement (0,41), l'estime de soi et les biais cognitifs (effets faibles), maintenus jusqu'à 1 an (Penney et al., *JAMA Psychiatry*, 2022) ; preuve débattue (section 14)
Thérapie MERIT (Lysaker & Klion, 2017)	Psychothérapie individuelle ciblant la réflexivité	Programme de recherche cohérent, essais encore limités (Lysaker et al., 2020)
Thérapie métacognitive (Wells)	Modifier les croyances sur ses propres pensées pour réduire rumination et inquiétude	Fondée sur le modèle S-REF (Wells, 2019)

8. Métacognition sociale et décision collective

La section 1 a posé, avec Frith, que la métacognition explicite s'est développée pour améliorer la décision collective. La recherche expérimentale permet de tester cette idée. Elle montre que la coopération repose sur l'échange de niveaux de confiance, et qu'elle se dégrade dès que ces niveaux ne sont pas calibrés de la même façon.

Deux têtes valent mieux qu'une, à condition de se ressembler

Bahrami et al. ont fait réaliser à des binômes une tâche de détection visuelle de bas niveau, en quatre expériences (*Science*, 2010). Chacun répondait d'abord seul; en cas de désaccord, les deux partenaires discutaient librement avant de trancher, sans jamais savoir s'ils avaient eu raison.

Les résultats dépendent de l'écart entre partenaires :

- **Entre deux personnes de sensibilité proche**, la décision commune est nettement meilleure que celle du meilleur des deux.

- **Quand l'un des deux voit nettement moins bien** — dans la quatrième expérience, l'image d'un partenaire était brouillée à son insu —, la décision commune devient pire que celle du meilleur seul (UCL, 2010).

Le modèle qui rend compte de ces deux résultats suppose que chacun communique fidèlement, à chaque essai, son niveau de confiance. La décision collective est donc une affaire de métacognition partagée : elle ne fonctionne que si chacun sait jauger sa propre fiabilité et l'exprimer dans une échelle que l'autre comprend.

On donne à chacun la même voix, à tort

Mahmoodi et al. ont répliqué ce dispositif au Danemark, en Iran et en Chine (*PNAS*, 2015). Pour décider au mieux, un groupe devrait pondérer les avis selon la compétence relative de chacun. Or les participants pondèrent presque également leur avis et celui de leur partenaire, quelle que soit la différence réelle de compétence.

Ce biais d'égalité résiste à tout : il persiste quand les participants reçoivent un retour explicite sur leur écart de compétence, et quand ils sont payés pour maximiser la justesse collective. Il coûte d'autant plus cher que l'écart entre les membres du groupe est grand. Les auteurs y voient un trait commun aux trois cultures : chacun se comporte comme si l'autre était aussi compétent que lui, en vertu d'une présomption d'égalité qui ne cède pas devant les preuves contraires (Royal Holloway, notice de l'article).

La surconfiance rapporte, la lucidité coûte

Anderson et al. ont cherché pourquoi la surconfiance est si répandue alors qu'elle nuit souvent à la performance (*JPSP*, 2012). Leurs six études aboutissent à une réponse sociale :

Études	Résultat
1 à 3	La surconfiance procure un statut social plus élevé, dans des groupes de courte comme de longue durée, en situation naturelle comme expérimentale
4	Elle produit une signature comportementale qui fait paraître compétent aux yeux des autres
5 et 6	Le désir de statut accroît, en retour, la surconfiance

Une étude ultérieure a vérifié ce qui se passe quand la surconfiance est démasquée : ses bénéfices de statut l'emportent encore sur ses coûts (Kennedy, Anderson & Moore, 2013).

Ce que ces trois résultats disent ensemble

Les groupes humains n'ont pas d'accès direct à la compétence de leurs membres : ils la lisent à travers la confiance exprimée. Cette lecture fonctionne bien entre personnes semblables, mal entre personnes différentes, et elle récompense l'assurance plus que la justesse. La métacognition n'est donc pas seulement une affaire privée : elle structure la manière dont les groupes décident et dont ils distribuent le crédit.

9. Expertise : quand la confiance des professionnels trompe

Les sections précédentes reposent surtout sur des tâches de laboratoire. Une objection naturelle est que les professionnels, formés et expérimentés, échappent à ces biais. Les données de terrain disent l'inverse : l'expertise améliore la performance, pas nécessairement le jugement que l'on porte sur elle.

Les médecins : une confiance qui ne suit pas la difficulté

Meyer et ses collègues ont soumis 118 médecins généralistes à quatre cas cliniques validés, deux faciles et deux difficiles (*JAMA Internal Medicine*, 2013). Le taux de

diagnostic juste tombait de 55,3 % sur les cas faciles à 5,8 % sur les cas difficiles. La confiance, notée sur 10, ne passait que de 7,2 à 6,4 (Medical Xpress, 2013).

Le résultat le plus préoccupant est le suivant : plus les médecins étaient confiants, moins ils demandaient d'examens complémentaires. Une confiance mal calibrée ne se contente pas de mal décrire la situation ; elle ferme la porte aux informations qui permettraient de la corriger. C'est exactement le circuit de contrôle décrit par Nelson et Narens en section 1 : une surveillance défaillante désactive le contrôle.

Les dirigeants financiers : des fourchettes trop étroites

Ben-David, Graham et Harvey ont exploité dix années d'enquêtes trimestrielles auprès de directeurs financiers américains, soit plus de 13 300 prévisions du rendement boursier assorties d'une fourchette à 80 % (NBER, 2010, publié dans le *Quarterly Journal of Economics* en 2013). Si ces fourchettes étaient bien calibrées, le rendement réalisé y tomberait huit fois sur dix. Il n'y tombe que 36 % du temps.

Deux prolongements donnent à ce chiffre sa portée :

- **La surconfiance se transfère.** Les dirigeants mal calibrés sur le marché le sont aussi sur les perspectives de leur propre entreprise.
- **Elle se traduit en décisions.** Les entreprises dont les dirigeants sont mal calibrés investissent davantage et recourent davantage à l'endettement.

La surconfiance n'est donc pas un trait de caractère sans conséquence : elle se lit dans les bilans.

Les investisseurs particuliers : trop sûrs, trop actifs

Barber et Odean ont analysé les comptes de plus de 35 000 foyers d'un courtier américain entre 1991 et 1997 (*Quarterly Journal of Economics*, 2001). Partant de

travaux montrant que les hommes se surestiment davantage dans les domaines financiers, ils comparent les comportements : les hommes échangent environ 45 % de plus que les femmes, et leur rendement net ajusté du risque est inférieur de 1,4 point par an. Chez les célibataires, l'écart atteint 67 % de transactions en plus et 2,3 points de rendement en moins.

Le mécanisme est métacognitif : qui surestime la précision de son information croit avoir raison plus souvent qu'il n'a raison, et paie en frais de transaction le prix de cette conviction.

La calibration s'apprend

Le résultat le plus encourageant vient des tournois de prévision géopolitique financés par l'agence de recherche du renseignement américain. L'équipe de Philip Tetlock et Barbara Mellers y a testé trois leviers : une formation aux probabilités, le travail en équipe, et le regroupement des meilleurs prévisionnistes en équipes d'élite (Mellers et al., *Psychological Science*, 2014). Les trois amélioraient à la fois la calibration et la capacité à distinguer les événements probables des improbables.

La formation elle-même durait moins d'une heure. Sur quatre années de tournoi, elle a amélioré la justesse des prévisions, mesurée par le score de Brier, d'environ 6 à 11 % par rapport à un groupe témoin, avec un effet significatif chaque année (Chang et al., *Judgment and Decision Making*, 2016, 11(5), 509-526). Apprendre à raisonner sur des classes de référence, à moyenner plusieurs estimations et à se méfier de ses propres certitudes ne transforme pas un amateur en expert; cela rend en revanche sa confiance plus fiable.

10. Intelligence artificielle : la métacognition déléguée

L'arrivée des outils d'intelligence artificielle dans le travail intellectuel déplace la question. Il ne s'agit plus seulement de juger sa propre pensée, mais de juger celle d'une machine, et de savoir quand s'y fier.

Le biais d'automatisation

Bien avant les modèles de langage, la recherche sur les systèmes d'aide à la décision médicale avait identifié le biais d'automatisation : la tendance à trop se reposer sur la machine. Une revue systématique de 74 études en recense les facteurs (Goddard, Roudsari et Wyatt, *JAMIA*, 2012) :

Famille de facteurs	Exemples
Utilisateur	Style cognitif, expérience du système, expérience de la tâche
Attitude	Confiance dans le système et confiance en soi
Environnement	Charge de travail, complexité de la tâche, pression du temps

Les remèdes identifiés sont de deux ordres : la formation et la responsabilisation de l'utilisateur d'un côté; de l'autre, la conception même du système, en affichant un niveau de confiance actualisé et en fournissant de l'information plutôt que des recommandations toutes faites.

Ce que les gens croient que l'IA sait

Steyvers et ses collègues ont mesuré l'écart entre la fiabilité réelle de grands modèles de langage et celle que leur prêtent les utilisateurs (*Nature Machine Intelligence*, 2025). Avec les explications fournies par défaut, les utilisateurs surestiment la justesse des réponses. Surtout, **des explications plus longues augmentent leur confiance sans que les réponses soient plus justes**. Lorsque les explications sont

ajustées pour exprimer l'incertitude interne du modèle, l'écart se réduit et les utilisateurs distinguent mieux les bonnes réponses des mauvaises (UC Irvine, 2025).

On retrouve le mécanisme de la section 6 : la fluidité d'un texte est prise pour un signe de maîtrise. Et celui de la section 8 : la confiance exprimée tient lieu de compétence. Une IA qui s'exprime avec assurance bénéficie de la même prime de statut que le participant surconfiant d'Anderson.

Conséquence : une compétence qui devient centrale

Ce dossier en tire une hypothèse, que les études citées rendent plausible sans la démontrer : à mesure que la production de textes, de calculs et d'analyses est déléguée, la valeur du travail humain se déplace vers le jugement porté sur cette production. Savoir quand vérifier, quand douter, quand demander une source, c'est de la métacognition appliquée à un tiers. Les organisations qui déploient ces outils sans former à cette vérification s'exposent au biais d'automatisation décrit en médecine depuis plus de dix ans.

11. Débat public : l'illusion de comprendre

La métacognition ne règle pas seulement l'apprentissage et la décision. Elle conditionne aussi la qualité du débat collectif, parce que chacun y défend des positions sur des sujets qu'il croit comprendre.

L'illusion de profondeur explicative

Rozenblit et Keil ont demandé à des participants d'évaluer leur compréhension d'objets courants, comme une fermeture éclair ou une chasse d'eau, puis d'en rédiger une explication causale détaillée, avant de réévaluer leur compréhension (*Cognitive Science*, 2002, 26, 521-562, dispositif rappelé par Fernbach et al., 2013). Après

l'exercice, les notes chutaient nettement : les participants découvraient qu'ils comprenaient beaucoup moins qu'ils ne le croyaient.

Expliquer modère, argumenter radicalise

Fernbach, Rogers, Fox et Sloman ont transposé le dispositif aux politiques publiques : relèvement de l'âge de la retraite, impôt à taux unique, rémunération des enseignants au mérite (*Psychological Science*, 2013). Trois expériences ont porté sur 198, 112 et 92 participants.

Consigne	Effet observé
Expliquer en détail comment la mesure produirait ses effets	La compréhension ressentie baisse pour les six politiques, et les positions deviennent moins extrêmes
Énumérer les raisons de sa position	Les positions ne se modèrent pas

Plus la compréhension ressentie baissait, plus les positions se modéraient. Les participants qui avaient adouci leur position après l'exercice étaient aussi moins enclins à donner de l'argent à une organisation militante. L'effet apparaissait chez les démocrates, les républicains et les indépendants (ScienceDaily, 2013).

La distinction entre les deux consignes est décisive. Donner ses raisons renforce la conviction; expliquer un mécanisme confronte à ses lacunes. Un débat public organisé autour de l'argumentation produit donc plus de certitude qu'un débat organisé autour de l'explication des effets attendus.

L'humilité intellectuelle

Leary et ses collègues ont mesuré un trait voisin, l'humilité intellectuelle, c'est-à-dire la conscience que ses propres croyances peuvent être fausses (*Personality and Social Psychology Bulletin*, 2017). Les personnes qui en ont davantage distinguent

mieux les arguments solides des arguments faibles, jugent moins sévèrement le caractère de ceux qui ne pensent pas comme elles, et acceptent mieux qu'un responsable politique change d'avis au vu de preuves nouvelles. Ce trait ne variait pratiquement pas entre progressistes et conservateurs, ni entre croyants et non-croyants.

12. Implications

Parce qu'elle règle à la fois l'apprentissage, la décision et la coopération, la métacognition a des conséquences pratiques dans plusieurs domaines. Leur niveau de preuve est inégal.

Domaine	Enjeu	Niveau de preuve
Éducation	Enseigner à surveiller et réguler son apprentissage	Élevé (section 6)
Santé	Savoir qu'on est malade et adhérer au soin	Hypothèse formulée par les chercheurs
Décision collective	Calibrer et échanger les niveaux de confiance	Expérimental (section 8)
Expertise et finance	Calibrer la confiance des décideurs, des praticiens et des prévisionnistes	Données de terrain et essais contrôlés (section 9)
Usage de l'IA	Juger la fiabilité d'une machine plutôt que s'y fier	Expérimental (section 10)
Débat public	Préférer l'explication des mécanismes à l'échange d'arguments	Expérimental (section 11)
Insertion sociale	Être jugé sur sa compétence plutôt que sur son assurance	Indirect, peu de données

Santé : la conscience de sa propre maladie

Présentant l'étude de 2010, Stephen Fleming prend l'exemple de deux patients, l'un conscient de sa maladie, l'autre non : le premier a plus de chances de suivre son traitement (Wellcome Trust, 2010). C'est une piste de recherche, pas un résultat établi par l'étude.

Organisations : corriger la lecture de la confiance

Les résultats de la section 8 désignent trois points d'action. Ce sont des déductions de ce dossier, non des dispositifs évalués :

- **Évaluer sur pièces, pas sur l'assurance** : la surconfiance procure du statut sans compétence supplémentaire. Épreuves en situation, critères écrits et décisions argumentées réduisent cette prime.
- **Pondérer les avis par la fiabilité constatée** : le biais d'égalité dégrade la décision collective. Garder trace de la justesse passée des avis permet de mieux les pondérer.
- **Rendre la confiance explicite et comparable** : la décision commune exige une échelle de confiance partagée. Demander à chacun d'exprimer sa certitude sur une même échelle rapproche les calibrages.

Insertion sociale : un coût de la différence

Les sections précédentes suggèrent un mécanisme par lequel une métacognition atypique peut gêner l'insertion :

1. **Le groupe juge sur la confiance affichée**, faute d'accès direct à la compétence.
2. **La personne lucide mais peu assurée** — le profil anxieux-dépressif — est perçue comme moins compétente qu'elle n'est.

3. **La personne mal calibrée**, dans un sens ou dans l'autre, dégrade la décision commune.
4. **Le groupe ne corrige pas spontanément** : le biais d'égalité résiste même à l'information.

Le seul lien direct documenté entre métacognition et emploi vient de la recherche clinique : dans la schizophrénie, une réflexivité faible s'accompagne de moins bonnes performances au travail (résultat de Lysaker et al., 2009, rappelé par Zalzala et al., *Journal of Nervous and Mental Disease*, 2022). Aucune statistique publique ne mesure l'emploi selon le profil métacognitif ; ce volet reste à documenter.

13. Objections et réponses

« **La métacognition n'est qu'un mot savant pour l'intelligence.** » Les données disent le contraire. La double dissociation de Rouault (section 5) montre que la métacognition varie avec des traits psychologiques alors que la performance n'en dépend pas, et que l'âge fait l'inverse. Fleming observe un substrat cérébral propre chez des participants de performance égale (section 3). On peut être compétent et mal se juger, ou moins compétent et lucide.

« **L'effet Dunning-Kruger est un artefact statistique, donc tout l'édifice tombe.** » La critique de Krueger et Mueller porte sur une analyse particulière, celle des quartiles de performance (section 14). Elle ne touche pas les mesures modernes, qui évaluent essai par essai la correspondance entre confiance et justesse et neutralisent l'effet de la performance (section 2). Les résultats de Rouault, Fleming ou Weil ne reposent pas sur la méthode contestée.

« **Ce sont des tâches de laboratoire sans rapport avec la vie réelle.** » C'est vrai pour une partie des travaux, et la section 14 le rappelle. Mais les données de la sec-

tion 9 sont des données de terrain : des diagnostics médicaux, plus de 13 000 prévisions de directeurs financiers, six années de transactions de 35 000 foyers. Elles montrent le même phénomène hors du laboratoire.

« **Faire douter les gens les rend indécis.** » La recherche ne vise pas le doute mais la calibration : être sûr quand on a raison, prudent quand on peut se tromper. Les prévisionnistes formés de Tetlock ne sont pas plus hésitants ; ils sont plus justes (section 9). Le profil anxieux-dépressif de Rouault montre d'ailleurs qu'une confiance basse n'est pas une bonne métacognition si elle reste basse quand on a raison.

« **Si la surconfiance rapporte du statut, pourquoi la corriger ?** » Parce qu'elle rapporte à l'individu ce qu'elle coûte au groupe. Les entreprises de dirigeants mal calibrés s'endettent davantage, les investisseurs surconfiants perdent du rendement, les décisions collectives se dégradent quand les avis sont mal pondérés (sections 8 et 9). Une organisation qui laisse la prime à l'assurance décider de ses promotions paie cette prime dans ses résultats.

14. Limites et controverses

Les résultats sont solides, mais six réserves doivent accompagner tout usage public.

1. **Des tâches de laboratoire.** Bahrami, Mahmoodi et Rouault utilisent des tâches perceptives simples. Leur transposition aux décisions de travail est plausible, pas démontrée.
2. **Des corrélations cérébrales sur de petits effectifs.** Le lien entre cortex préfrontal antérieur et métacognition repose sur une étude de 32 participants ; il indique une association, pas une cause, et ne dit rien du caractère inné ou acquis de cette structure.

3. **L'effet Dunning-Kruger est discuté.** Une partie de la littérature attribue le résultat à un effet « meilleur que la moyenne » combiné à un artefact de régression statistique, plutôt qu'à un déficit métacognitif (Krueger & Mueller, *JPSP*, 2002) : en neutralisant statistiquement l'un ou l'autre de ces deux facteurs, l'asymétrie disparaît.
4. **La preuve sur le MCT est disputée.** Une méta-analyse de 2016 trouve des effets faibles et non significatifs, ramenés à 0,03 sur les délires après correction des biais de publication (van Oosterhout et al., *Psychological Medicine*, 2016). Une autre, la même année, conclut à un effet faible à modéré (Eichner & Berna, *Schizophrenia Bulletin*, 2016).
5. **Corrélation n'est pas causalité.** Dans la schizophrénie, le lien entre réflexivité et emploi est prédictif; symptômes, neurocognition et stigmatisation pèsent aussi.
6. **Aucune donnée d'insertion par profil métacognitif.** Le dossier ne peut pas chiffrer le nombre de personnes concernées ni leur taux d'emploi.

SOURCES

Chaque référence a été contrôlée en ligne (notice, résumé ou texte intégral) en octobre puis en novembre 2026.

Fondements et mesure

- Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist*, 34(10), 906–911.
- Nelson, T. O. & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation*, 26, 125–173.
- Fleming, S. M. & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, 443.

- Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*.

Cerveau et développement

- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541–1543.
- Goupil, L., Romand-Monnier, M. & Kouider, S. (2016). Infants ask for help when they know they don't know. *PNAS*, 113(13), 3492–3496.
- Weil, L. G. et al. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, 22(1), 264–271.
- Goupil, L. & Kouider, S. (2016). Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Current Biology*, 26(22), 3038–3045.
- Moses-Payne, M. E. et al. (2021). I know better! Emerging metacognition allows adolescents to ignore false advice. *Developmental Science*.
- Smith, J. D., Shields, W. E. & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26(3), 317–339.

Différences individuelles et apprentissage

- Krueger, J. & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2), 180–188.
- Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, 84(6), 443–451.
- Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories: is spacing the « enemy of induction »? *Psychological Science*, 19(6), 585–592.
- Kornell, N., Castel, A. D., Eich, T. S. & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, 25(2), 498–503.
- Education Endowment Foundation. Metacognition and self-regulation, *Teaching and Learning Toolkit* (consulté en novembre 2026).
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, 14(1), 4–58.

Santé mentale

- Wells, A. (2019). Breaking the cybernetic code: understanding and treating the human metacognitive control system. *Frontiers in Psychology*, 10, 2621.
- Lysaker, P. H. et al. (2020). Schizophrenia, recovery and the self: an introduction to the special issue on metacognition. *Schizophrenia Research: Cognition*, 19, 100167.
- Kukla, M. & Lysaker, P. H. (2020). Metacognition over time is related to neurocognition, social cognition, and intrapsychic foundations in psychosis. *Schizophrenia Research: Cognition*, 19, 100149.
- Penney, D. et al. (2022). Immediate and sustained outcomes and moderators associated with metacognitive training for psychosis. *JAMA Psychiatry*.
- Eichner, C. & Berna, F. (2016). Acceptance and efficacy of metacognitive training on positive symptoms and delusions. *Schizophrenia Bulletin*, 42(4), 952–962.
- Zalzala, A. et al. (2022). Metacognitive Training (MCT) to improve insight and work outcome in schizophrenia. *The Journal of Nervous and Mental Disease*, 210(9), 655–658.
- van Oosterhout, B. et al. (2016). Metacognitive training for schizophrenia spectrum patients: a meta-analysis on outcome studies. *Psychological Medicine*, 46, 47–57.

Dimension sociale

- Frith, C. D. (2012). The role of metacognition in human social interactions. *Phil. Trans. R. Soc. B*, 367(1599), 2213–2223.
- Bahrami, B. et al. (2010). Optimally interacting minds. *Science*, 329(5995), 1081–1085.
- Mahmoodi, A. et al. (2015). Equality bias impairs collective decision-making across cultures. *PNAS*, 112(12), 3835–3840.
- Anderson, C., Brion, S., Moore, D. A. & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, 103(4), 718–735.
- Kennedy, J. A., Anderson, C. & Moore, D. A. (2013). When overconfidence is revealed to others: testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, 122(2), 266–279.

Expertise, intelligence artificielle et débat public

- Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R. & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, 173(21), 1952–1958.
- Ben-David, I., Graham, J. R. & Harvey, C. R. (2013). Managerial miscalibration. *Quarterly Journal of Economics*, 128(4), 1547–1584 (document de travail NBER 16215, 2010).

- Barber, B. M. & Odean, T. (2001). Boys will be boys: gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, 116(1), 261–292.
- Mellers, B. et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*.
- Chang, W., Chen, E., Mellers, B. & Tetlock, P. (2016). Developing expert political judgment: the impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5), 509–526.
- Goddard, K., Roudsari, A. & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *JAMIA*, 19(1), 121–127.
- Steyvers, M. et al. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*.
- Rozenblit, L. & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Fernbach, P. M., Rogers, T., Fox, C. R. & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*.
- Leary, M. R. et al. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6), 793–813.