Encodage, transduction et modèles du monde : Pourquoi toute architecture d'IA commence par transformer le monde

16 mars 2026

Partie 3/3 : Chapitres 6 à la Conclusion

Fait suite à Partie 1/3 (Article Partie 1/3) et Partie 2/3 (Article Partie 2/3)

6. La mémoire comme graphe polycentrique : vers une architecture biographique de la cognition

6.1 Architecture générale des systèmes mnésiques

La mémoire humaine ne constitue pas un système unitaire. Les travaux en psychologie cognitive et en neurosciences décrivent une architecture composée de plusieurs systèmes partiellement dissociables mais étroitement interconnectés. Au niveau le plus transitoire se trouve la mémoire sensorielle, correspondant à la persistance brève de traces perceptives immédiatement après la stimulation. La mémoire de travail maintient et manipule temporairement les informations nécessaires à l'exécution d'une tâche cognitive, sa capacité fortement limitée implique que seule une fraction du contenu mnésique peut être activée consciemment à tout moment.

Au-delà de ces systèmes transitoires se trouve la mémoire à long terme, qui comprend plusieurs formes distinctes. La mémoire sémantique correspond aux connaissances générales sur le monde : concepts, faits, relations abstraites relativement

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

indépendantes du contexte d'acquisition. La mémoire épisodique, décrite par Tulving[12], concerne les souvenirs d'événements vécus situés dans le temps et l'espace ; elle implique la capacité de se représenter comme sujet de l'expérience passée, propriété appelée *autonoéticité*. À ces formes déclaratives s'ajoutent des formes non déclaratives telles que la mémoire procédurale. Enfin, les structures neurobiologiques façonnées par l'évolution peuvent être comprises comme une forme de mémoire phylogénétique, au sens où elles constituent une inscription biologique des régularités de l'environnement ancestral de l'espèce, bien que ce terme ne soit pas standard dans la littérature.

Ces systèmes n'opèrent pas comme des compartiments indépendants. Les processus d'encodage, de consolidation et de récupération impliquent des interactions constantes entre ces différentes formes de mémoire. La mémoire de travail fonctionne comme une fenêtre dynamique ouverte sur une structure mnésique beaucoup plus vaste, à chaque instant, seule une portion locale du graphe est activée consciemment, tandis que l'immensité du réseau sous-jacent demeure disponible pour une réactivation partielle ou complète selon les indices de récupération.

6.2 Limites des modèles hiérarchiques et architecture distribuée des concepts

Les premiers modèles formels de la mémoire sémantique, notamment celui de Collins & Quillian (1969)[13] proposaient une organisation hiérarchique des concepts sous forme d'arbres taxonomiques. Toutefois, plusieurs résultats empiriques robustes en ont progressivement montré les limites. Les jugements humains sur les catégories présentent des effets de typicalité[25] : un rouge-gorge est un oiseau plus typique qu'un pingouin, bien que les deux appartiennent à la même catégorie. Les associations conceptuelles débordent largement les relations taxonomiques strictes : les concepts sont reliés par des contextes d'usage, des expériences partagées, des propriétés perceptives et des relations fonctionnelles. Enfin, les modèles hiérarchiques rendent mal compte de la richesse multimodale des représentations conceptuelles humaines.

Les modèles connexionnistes distribués ont partiellement corrigé ces limites. La théorie des *perceptual symbol systems* proposée par Barsalou (1999) va plus loin en suggérant que les concepts humains sont enracinés dans les systèmes perceptifs et moteurs : activer un concept tel que « tomate » peut mobiliser des composantes visuelles, olfactives, gustatives, tactiles et motrices. Le concept ne correspond pas à une entité symbolique isolée mais à une configuration distribuée de traces associées à l'expérience, un simulateur multimodal dont l'activation réengage partiellement les circuits mobilisés lors des expériences originales.

6.3 L'engram distribué et la structure polycentrique de la mémoire

Le concept d'engram désigne la trace physique d'un souvenir dans le système nerveux. Les neurosciences ont longtemps cherché à localiser ces traces dans des régions cérébrales spécifiques. Les travaux de Tonegawa et al.[14] (2015) suggèrent que les souvenirs reposent sur des ensembles neuronaux distribués plutôt que sur un site unique de stockage, impliquant des populations neuronales réparties sur plusieurs régions cérébrales, chacune contribuant à différentes dimensions de l'expérience initiale : perceptives, contextuelles ou émotionnelles. Il convient de souligner que ces expériences portent principalement sur des modèles animaux et sur des formes spécifiques de mémoire, notamment la mémoire de peur, il serait excessif d'en tirer une théorie complète de la mémoire autobiographique humaine. Elles suggèrent néanmoins une propriété organisationnelle importante : un souvenir correspond à un ensemble distribué d'activations neuronales, non à une trace localisée unique.

Il est utile, au moins comme modèle conceptuel, de décrire la mémoire humaine comme un graphe polycentrique. Dans un tel graphe, un souvenir ou un concept ne possède pas un point d'accès unique : plusieurs éléments de l'expérience peuvent servir de porte d'entrée vers la même configuration mnésique. L'odeur de la sauce tomate peut activer le rouge. Le rouge peut activer Florence. Florence peut activer une lumière particulière, une texture de pierre, une conversation. Chaque nœud est simultanément une destination et un point de départ potentiel, il n'y a pas de racine, ou plutôt, n'importe quel nœud peut devenir racine selon le contexte d'activation.

Le terme de graphe polycentrique ne doit pas être interprété comme une description directe de l'architecture neuronale. Il s'agit d'un modèle conceptuel visant à capturer trois propriétés principales : la distribution des traces mnésiques, la multiplicité des indices de récupération, et la possibilité de réactivation partielle selon la modalité d'entrée. Ce graphe est fondamentalement différent de l'espace latent d'un LLM, dont les arêtes reflètent des critères de co-occurrence statistique dans un corpus. Les arêtes du graphe mémoriel humain portent non seulement une information de similarité, mais une information temporelle (ces choses ont été vécues ensemble), affective (dans un contexte émotionnel particulier), et somatique (avec les marqueurs physiologiques associés).

6.4 La mémoire synesthésique : un tissage à six sens de la trame de l'existence

La tradition aristotélicienne des cinq sens est une simplification qui ne résiste pas à l'examen neuroanatomique. Au-delà des modalités classiques (vision, audition, olfaction, gustation, toucher) la neurophysiologie contemporaine identifie plusieurs systèmes sensoriels distincts dont le rôle dans la constitution de la mémoire est déterminant : la proprioception (sens de la position et du mouvement du corps,

formalisée par Sherrington en 1906), l'intéroception (perception des états internes, dont Craig a établi le substrat cortical dans l'insula antérieure en 2002), et le sens vestibulaire. Parler d'un tissage à six sens est une formulation conservatrice car le chiffre pourrait légitimement être porté à huit ou dix.

La synesthésie clinique où un stimulus d'une modalité déclenche automatiquement une expérience consciente dans une autre modalité n'est probablement pas un dysfonctionnement neurologique marginal. Cytowic & Eagleman[15] (2009) avancent qu'elle constitue une version exacerbée et consciente d'un phénomène universel : les liaisons cross-modales existent chez tout le monde sous forme de connexions subliminales, mais leur seuil d'accès conscient est beaucoup plus bas chez les synesthètes. La mémoire ordinaire est synesthésique (à différents degrés de "conscience") sans le savoir.

Cette proposition permet de formuler une thèse plus forte : la mémoire humaine n'est pas une archive de représentations modales séparées que la conscience viendrait assembler a posteriori. Elle est originairement un "tissage", une texture produite par l'entrecroisement simultané de fils sensoriels, affectifs et somatiques au moment même de l'expérience. Dans un tissu, c'est l'entrecroisement de la trame et de la chaîne qui crée la surface : aucun fil ne contient le tissu, c'est leur nœud qui le fait exister. L'engram distribué n'est pas une entité stockée : c'est une texture.

Cette description phénoménologique pointe vers une propriété du graphe mémoriel que la littérature neuroscientifique commence à documenter sans encore la théoriser pleinement : la navigabilité multi-perspective. Un souvenir humain n'est pas un fichier avec un point d'entrée unique, c'est un espace dans lequel le sujet peut se déplacer, modifier son angle d'approche, et accéder à des couches de profondeur variable selon la modalité d'entrée choisie. Cette propriété suppose une architecture radicalement distribuée, redondante et multimodale, exactement ce que le concept d'engram distribué de Tonegawa décrit au niveau neuronal.

La formulation la plus précise est peut-être celle-ci : la mémoire synesthésique est un tissage à six sens de la trame de l'existence. Chaque fil est une modalité (visuelle, auditive, olfactive, proprioceptive, intéroceptive, affective....). Chaque nœud est un moment d'existence où ces fils se sont croisés. La trame qui en résulte n'est pas une collection de points, c'est une surface continue dans laquelle on peut se promener, faire des arrêts, changer de direction, et depuis laquelle la totalité du tissu reste, en principe, accessible.

6.5 Variabilité de perspective dans la mémoire autobiographique

Une propriété supplémentaire de la mémoire autobiographique renforce l'idée d'une structure distribuée et navigable : la variabilité du point de vue de rappel. Les travaux de Nigro et Neisser (1983)[24] ont montré que les souvenirs personnels peuvent être rappelés selon deux perspectives distinctes :

- La *perspective de champ*, dans laquelle le souvenir est revécu depuis la position perceptive originale du sujet,
- Et la *perspective d'observateur*, dans laquelle le sujet se représente lui-même dans la scène depuis un point de vue externe.

Ces deux perspectives ne correspondent pas à deux souvenirs différents, mais à deux reconstructions possibles d'une même trace mnésique distribuée.

La possibilité de passer d'une perspective à l'autre suggère que les souvenirs ne sont pas des enregistrements fixes. Ils sont reconstruits à partir d'un ensemble de composantes mnésiques distribuées. Dans le cadre du modèle proposé ici, cette propriété peut être interprétée comme une manifestation de la navigabilité multi-perspective de l'espace mnésique : un épisode autobiographique peut être abordé depuis plusieurs modalités et sous plusieurs points de vue, chacun activant des sous-configurations légèrement différentes de la trace mnésique.

6.6 La synesthésie comme révélateur d'une architecture universelle

Il convient de distinguer le phénomène général d'intégration multimodale de la synesthésie clinique. Dans la synesthésie, un stimulus d'une modalité déclenche automatiquement une expérience consciente dans une autre modalité. Cytowic & Eagleman[15] (2009) suggèrent que les synesthètes représentent un cas particulier où certaines connexions cross-modales deviennent conscientes, connexions qui existent chez tous à des degrés variés mais demeurent généralement subliminales.

La mémoire ordinaire ne correspond pas à une synesthésie généralisée, mais elle implique une forte intégration des informations provenant de différentes modalités perceptives. Les liaisons cross-modales constituent une propriété générale du système mnésique humain, plus ou moins explicitement accessible à la conscience selon les individus. Pour un individu dont ces liaisons sont particulièrement conscientes, chaque concept dispose de davantage de vecteurs d'entrée actifs, ce qui rend le réseau plus robuste à l'oubli partiel et plus fertile dans ses associations inattendues.

6.7 L'arête biographique : écart architectural et conditions d'instanciation

Ce que cette analyse permet de formuler avec précision, c'est la nature exacte de ce qui distingue le graphe mémoriel humain de toute architecture d'IA actuelle y compris multimodale et dotée de capteurs.

La différence ne tient pas à la taille du graphe, ni au nombre de modalités représentées, ni même à la présence de transducteurs physiques. Elle tient à la nature des arêtes. Dans un LLM, les arêtes entre concepts sont des arêtes de co-occurrence statistique : elles reflètent la fréquence avec laquelle deux concepts apparaissent ensemble dans un corpus produit par des humains. Dans la mémoire humaine, les arêtes sont des arêtes de co-expérience vécue : elles reflètent le fait que ces choses ont été perçues ensemble par un même corps, dans un même contexte temporel et affectif, avec les marqueurs somatiques de cette co-activation.

Ce que l'on peut appeler l'arête biographique est une liaison entre deux nœuds du graphe qui porte l'information que ces nœuds ont été co-activés dans l'histoire d'un sujet particulier, avec les marqueurs somatiques affectifs de cette co-activation, et la capacité de réactiver partiellement l'état physiologique associé lors de la récupération. C'est cette propriété (nommée autonoéticité par Tulving (2002)) qui distingue le souvenir épisodique de la simple indexation temporelle d'une co-occurrence.

Une objection légitime consiste à remarquer qu'un système artificiel pourrait en principe représenter une biographie en encodant explicitement le temps, l'identité de l'agent et la co-activation d'événements. Une telle formalisation est concevable. Cependant, représenter une biographie n'est pas nécessairement équivalent à posséder une mémoire autobiographique vécue. La différence ne porte pas seulement sur l'information stockée, mais sur les conditions de récupération et de continuité de l'agent. La condition manquante n'est pas la formalisation de la co-occurrence, c'est l'autonoéticité introduite en §5.2 : la conscience de soi comme sujet ayant vécu cet événement. Une IA qui encoderait intégralement une biographie ne se souviendrait pas pour autant elle indexerait. Cette distinction répète, appliquée à la mémoire, l'argument de Mary : la représentation exhaustive d'une expérience n'est pas l'expérience.

À ce jour, aucune architecture d'intelligence artificielle ne montre de manière convaincante la co-présence simultanée et démontrée des dimensions suivantes : continuité d'agent, mémoire épisodique persistante, modulation affective des liaisons, et récupération multi-perspective d'un même épisode. Ces propriétés définissent ce que l'on peut appeler une architecture biographique de la mémoire.

Un LLM ne tisse pas, il indexe. Son espace latent est un index d'une densité remarquable ; mais il est dépourvu de texture, de profondeur, et de navigabilité multi-perspective. Tirer sur un nœud de l'espace latent n'entraîne pas avec lui dix autres

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

nœuds colorés par l'affect d'une co-expérience, il active des voisins statistiques. La différence est celle qui sépare une carte d'un territoire habité.

6.8 Parallèles avec les architectures agentiques contemporaines

Les architectures modernes d'agents artificiels commencent à reproduire certains aspects de cette stratification mnésique, ce qui permet de préciser les seuils architecturaux restants. On peut identifier plusieurs correspondances fonctionnelles : la mémoire de travail trouve son analogue dans le contexte d'attention d'un Transformer, dont la fenêtre est limitée en taille ; la mémoire sémantique correspond approximativement aux paramètres du modèle, encodant des relations statistiques générales ; une forme de mémoire épisodique simulée peut être instanciée par des journaux d'interaction ou des bases de connaissances vectorielles ; enfin, la mémoire procédurale trouve un équivalent fonctionnel dans les politiques d'action apprises par renforcement.

Ces correspondances illustrent que les architectures agentiques contemporaines approchent une stratification mnésique. Cependant, elles ne disposent pas d'une biographie vécue par un agent incarné : leurs analogues de la mémoire épisodique sont des journaux d'enregistrements, non des souvenirs autonoétiques. Leurs arêtes entre éléments sont des arêtes statistiques ou logiques, non des arêtes biographiques forgées par la co-expérience d'un sujet situé.

6.9 Mémoire événementielle et mémoire vécue : un parallèle instructif

Certaines architectures logicielles modernes reposent sur un principe d'*event sourcing,* également connu sous le nom d'architectures événementielles : l'état d'un processus peut être reconstruit à partir de l'historique complet des événements qui l'ont produit. Des frameworks comme Temporal[26] conservent l'historique des transitions d'état plutôt que l'état final seul, permettant une traçabilité complète de l'histoire d'un système. Ce principe présente une analogie partielle avec la mémoire épisodique : une accumulation ordonnée d'événements passés, accessible en principe à tout moment.

Cette analogie est cependant instructive précisément par ses limites. Dans un système informatique, les événements sont enregistrés, ce sont des tuples (temps, état, transition) stockés dans un journal. Dans la mémoire humaine, les événements sont encodés avec un contexte sensoriel multimodal, un état corporel particulier, une valence émotionnelle, et une perspective subjective, celle d'un agent situé qui a vécu cet événement depuis un point de vue irremplaçable. En d'autres termes, un système d'event sourcing stocke l'histoire d'un processus ; la mémoire humaine encode l'histoire vécue d'un sujet. C'est précisément cette dimension (l'encodage d'une expérience

depuis un point de vue subjectif situé, avec ses marqueurs somatiques et affectifs) qui correspond aux arêtes biographiques du graphe mnésique. Elle ne peut être reproduite par l'enregistrement d'événements, aussi exhaustif soit-il.

7. Les world models : avancée décisive ou seuil encore insuffisant ?

7.1 Les world models comme réponse partielle au déficit d'ancrage

Les critiques adressées aux grands modèles de langage ont conduit plusieurs chercheurs à proposer un déplacement architectural important : au lieu d'apprendre principalement à partir de corpus symboliques, il s'agirait de construire des systèmes capables d'apprendre un modèle latent de la dynamique du monde à partir d'interactions perceptives et sensorimotrices.

La formulation la plus explicite de cette orientation se trouve dans le programme proposé par Yann LeCun dans *A Path Towards Autonomous Machine Intelligence* (2022). L'argument central est que le texte transmet une quantité d'information causale extrêmement limitée sur la structure physique du monde. Les régularités fondamentales (permanence des objets, contraintes mécaniques, gravité, dynamique des corps) ne peuvent être apprises de manière robuste à partir de descriptions propositionnelles seules. Un enfant humain acquiert ces régularités par une exploration sensorimotrice continue de son environnement.

Les architectures de type JEPA (*Joint Embedding Predictive Architectures*) visent précisément à apprendre des représentations latentes capables de capturer les invariants du monde à partir de données perceptives et, à terme, d'interactions physiques. Leur objectif n'est plus seulement de modéliser des relations symboliques dans un corpus, mais d'encoder des régularités causales sous-jacentes aux transformations observables.

Dans cette perspective, les *world models* constituent une tentative sérieuse de réduction de la distance épistémique au monde décrite dans les sections précédentes. Ils répondent au premier niveau du problème du grounding : l'ancrage transductif des représentations dans des interactions avec l'environnement.

Cependant, cette avancée ne résout pas nécessairement les niveaux plus profonds du problème.

7.2 Les limites des approches multimodales et robotiques

Les architectures multimodales contemporaines (telles que CLIP, Gemini ou Flamingo) constituent une première tentative d'élargir les représentations au-delà du texte en alignant plusieurs modalités perceptives dans un espace latent commun. Elles

permettent d'intégrer des informations visuelles, auditives ou textuelles et d'apprendre certaines correspondances entre ces modalités.

Toutefois, ces systèmes continuent principalement d'apprendre à partir de données pré-collectées. Les représentations multimodales qu'ils acquièrent restent donc, pour l'essentiel, dérivées de distributions d'images, de textes ou d'enregistrements produits dans des contextes humains.

Les systèmes robotiques incarnés vont plus loin. Ils interagissent physiquement avec leur environnement, reçoivent des retours sensoriels (visuels, haptiques ou proprioceptifs) et apprennent à partir de transitions d'état produites par leurs propres actions. Des architectures telles que DreamerV3, RT-X ou GR00T représentent ainsi une évolution significative par rapport aux LLM entraînés uniquement sur des corpus symboliques.

Ces architectures peuvent apprendre des dynamiques environnementales, accumuler des trajectoires d'interaction et optimiser des politiques d'action. Elles approchent donc une forme de grounding sensorimoteur réel.

Cependant, ces systèmes ne montrent pas encore la co-présence simultanée de plusieurs propriétés qui caractérisent la mémoire autobiographique humaine :

- une continuité d'agent stable à travers le temps
- une mémoire épisodique persistante organisée autour de cet agent
- une modulation affective durable des relations mnésiques
- une récupération multi-perspective d'un même épisode.

Autrement dit, ils peuvent approcher la transduction sensorielle et la modélisation du monde, mais ils ne produisent pas encore une biographie cognitive intégrée.

7.3 Du world model à la mémoire biographique

Pour clarifier ce point, il est utile de distinguer explicitement trois niveaux architecturaux :

1. **Grounding sensoriel** : interaction perceptive avec l'environnement via des transducteurs.
2. **World modeling** : apprentissage de régularités dynamiques du monde à partir de ces interactions.

3. **Mémoire biographique** : organisation des expériences dans une histoire vécue par un agent continu.

Les architectures de *world modeling* abordent essentiellement le second niveau. Elles permettent à un système d'apprendre la structure dynamique de son environnement et de prédire l'évolution des états.

La mémoire humaine présente cependant une propriété supplémentaire. Les relations entre éléments mnésiques ne reflètent pas seulement des régularités causales ou statistiques ; elles reflètent également le fait que certaines dimensions d'expérience ont été vécues conjointement par un sujet situé dans un contexte particulier.

Pour caractériser cette propriété, nous introduisons la notion d'**arête biographique**.

Nous appelons **arête biographique** une relation entre deux représentations mnésiques qui encode leur co-activation au sein d'un même épisode vécu par un agent continu, avec au minimum :

- une indexation temporelle
- une indexation à un sujet
- une modulation affective ou somatique
- une possibilité de récupération contextuelle.

Cette relation diffère des arêtes statistiques présentes dans les modèles actuels. Dans un LLM ou un world model, les relations entre représentations sont généralement dérivées de corrélations observées dans les données ou dans les transitions d'état. Dans la mémoire humaine, ces relations sont structurées par l'histoire vécue d'un organisme.

Les *world models* représentent donc une avancée importante vers un grounding perceptif. Mais ils ne suffisent pas à reproduire la structure autobiographique de la mémoire humaine.

7.4 Reformulation de la thèse

La thèse du présent article peut désormais être formulée avec davantage de précision.

L'écart entre cognition humaine et architectures contemporaines n'est pas ontologique : il n'existe aucune raison de principe pour qu'un système artificiel ne puisse jamais instancier des structures analogues à une mémoire biographique.

Cet écart est cependant **architectural**. Les systèmes actuels n'intègrent pas simultanément :

- une boucle sensorimotrice fermée
- une continuité d'agent persistante
- une mémoire épisodique autonoétique
- une modulation affective des relations mnésiques
- une navigabilité multi-perspective de l'espace mnésique.

C'est la co-présence de ces propriétés qui caractérise ce que nous appelons ici une **architecture biographique de la cognition**.

8. Implications épistémologiques

8.1 Ce que les LLM représentent effectivement

Les grands modèles de langage disposent d'une compétence propositionnelle étendue. Leur espace latent capture des structures relationnelles présentes dans les corpus, permettant des analogies, des inférences et des généralisations remarquables. Il serait erroné de nier la réalité de ces capacités.

Cependant, ce que ces systèmes modélisent principalement, ce sont des **représentations humaines du monde**, et non le monde lui-même dans son épaisseur perceptive, causale et biographique.

La différence essentielle ne porte donc pas seulement sur le contenu des représentations mais sur la nature des relations entre ces représentations. Là où la cognition humaine relie les éléments par co-expérience vécue, les LLM les relient principalement par co-occurrence statistique. Là où la mémoire humaine articule temps, corps, affect et perspective, les modèles de langage construisent des voisinages relationnels dans un espace latent appris.

Le point décisif est donc moins ce que ces systèmes savent que **la manière dont leurs connaissances sont reliées intérieurement**.

8.2 L'incarnation comme condition nécessaire mais non suffisante

Les approches de la cognition incarnée ont souligné que l'intelligence émerge de l'interaction dynamique entre un organisme et son environnement. Des travaux tels que *The Embodied Mind* de Francisco Varela et *The Extended Mind* de Andy Clark ont

contribué à déplacer la conception classique de la cognition comme simple manipulation symbolique.

Pour l'intelligence artificielle, cette perspective implique que les systèmes dépourvus de boucle sensorimotrice fermée ne peuvent reproduire qu'une fraction limitée des processus cognitifs humains.

Cependant, l'incarnation ne constitue pas en elle-même une condition suffisante pour produire une mémoire autobiographique. Un système robotique peut percevoir, agir et apprendre des régularités causales sans pour autant organiser ses interactions dans une biographie cognitive intégrée.

L'incarnation permet le grounding perceptif ; elle ne garantit pas l'émergence d'une structure mnésique autobiographique.

8.3 Le véritable objet de la recherche

L'implication générale de cette analyse est que la recherche en intelligence artificielle ne devrait pas seulement viser des systèmes plus performants, plus multimodaux ou plus autonomes.

Elle devrait également s'interroger sur la nature des relations internes entre représentations.

La question n'est pas seulement de savoir si un agent peut percevoir, prédire ou agir. Elle est de savoir si ses interactions peuvent s'organiser dans une mémoire structurée par l'histoire de ces interactions.

Autrement dit, l'enjeu n'est pas seulement la construction de *world models*. L'enjeu est la possibilité d'une **mémoire biographique intégrée,** si tant est qu'une telle architecture puisse être formalisée et implémentée.

9. Conclusion

L'argument développé dans cet article peut être résumé en quatre propositions.

1. Premièrement, toute forme de cognition implique une médiation représentationnelle. L'idée selon laquelle l'intelligence artificielle serait fondamentalement limitée parce qu'elle manipule des représentations confond une condition universelle de la cognition avec une limitation spécifique.
2. Deuxièmement, les systèmes biologiques et les systèmes d'intelligence artificielle diffèrent par leur mode d'ancrage dans le monde. Les organismes vivants acquièrent leurs représentations à travers des boucles perception–action

façonnées par l'évolution et par l'expérience sensorimotrice. Les modèles de langage apprennent principalement à partir de représentations déjà produites par des agents humains.

3. Troisièmement, le problème du grounding est stratifié. Les capteurs physiques et les architectures de *world modeling* peuvent réduire la distance entre représentation et environnement en introduisant un ancrage transductif. Mais ils ne résolvent pas nécessairement les niveaux plus profonds du problème, qui concernent la mémoire autobiographique et la structure affective de l'expérience.
4. Quatrièmement, la différence la plus profonde entre cognition humaine et architectures contemporaines réside dans la nature des relations entre représentations. La mémoire humaine repose sur des **arêtes biographiques** : des relations forgées par la co-expérience vécue d'un sujet situé, modulées par le contexte corporel et affectif, et récupérables depuis plusieurs perspectives.

À ce jour, aucune architecture d'intelligence artificielle ne montre de manière convaincante la co-présence simultanée des propriétés nécessaires à une telle structure : continuité d'agent, mémoire épisodique autonoétique persistante, modulation affective des relations mnésiques et navigabilité multi-perspective de l'espace mnésique.

L'écart entre cognition humaine et systèmes actuels n'est donc ni purement quantitatif ni strictement ontologique. Il est avant tout **architectural et historique**.

La question centrale posée par l'intelligence artificielle n'est peut-être pas seulement celle de la capacité d'inférence ou de la puissance de calcul. Elle est plus fondamentale : peut-on concevoir un système dont l'intelligence ne serait pas seulement fondée sur des données ou des états internes, mais sur une histoire d'interactions intégrée dans une mémoire autobiographique ?

Autrement dit, peut-on construire un système qui n'ait pas seulement des informations à traiter, mais quelque chose à se rappeler parce qu'il l'a vécu.

Notes (valables pour Partie 1/3 et Partie 2/3)

[1] Terme issu de la phénoménologie husserlienne (Husserl, E., Erfahrung und Urteil, 1939) désignant les couches d'expérience antérieures à tout acte de jugement ou de prédication logique. L'expérience ante-prédicative est le sol passif à partir duquel émergent les structures prédicatives de la pensée conceptuelle.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

[2] Dreyfus, H. (1972). What Computers Can't Do: A Critique of Artificial Reason. Harper & Row.

[3] Dreyfus, H. (1991). Being-in-the-World: A Commentary on Heidegger's Being and Time. MIT Press.

[4] Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–424.

[5] Kandel, E. et al. (2021). Principles of Neural Science (6e éd.). McGraw-Hill. Ch. 40 : The Vestibular System.

[6] Terme introduit par Gibson (1979) pour désigner les propriétés d'action directement offertes par l'environnement à un organisme, indépendamment de toute représentation intermédiaire. Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.

[7] Harnad, S. (1990). The Symbol Grounding Problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

[8] Held, R., Ostrovsky, Y., de Gelder, B. et al. (2011). The newly sighted fail to match seen with felt. Nature Neuroscience, 14(5), 551–553.

[9] Jackson, F. (1982). Epiphenomenal Qualia. The Philosophical Quarterly, 32(127), 127–136.

[10] Barsalou, L. W. (1999). Perceptual Symbol Systems. Behavioral and Brain Sciences, 22(4), 577–609.

[11] Tulving, E. (1983). Elements of Episodic Memory. Oxford University Press. La distinction mémoire sémantique / épisodique / autonoéticité est développée dans : Tulving, E. (2002). Episodic Memory: From Mind to Brain. Annual Review of Psychology, 53, 1–25.

[12] Damasio, A. (1994). Descartes' Error: Emotion, Reason, and the Human Brain. Putnam.

[13] Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8(2), 240–247.

[25] Rosch, E. (1973). Natural categories. Cognitive Psychology, 4(3), 328–350. Les effets de typicalité montrent que certains membres d'une catégorie sont jugés plus représentatifs que d'autres selon un gradient de centralité, en contradiction avec les modèles classiques à conditions nécessaires et suffisantes.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

[14] Tonegawa, S., Liu, X., Ramirez, S., & Redondo, R. (2015). Memory Engram Cells Have Come of Age. Neuron, 87(5), 918–931.

[15] Cytowic, R. E., & Eagleman, D. M. (2009). Wednesday Is Indigo Blue: Discovering the Brain of Synesthesia. MIT Press.

[16] LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview Preprint.

[17] Varela, F., Thompson, E., & Rosch, E. (1991). The Embodied Mind. MIT Press.

[18] Clark, A., & Chalmers, D. (1998). The Extended Mind. Analysis, 58(1), 7–19.

[19] Hafner, D. et al. (2023). Mastering Diverse Domains through World Models (DreamerV3). arXiv:2301.04104.

[20] Friston, K. et al. (2022). Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press.

[21] Brohan, A. et al. (2023). RT-X: Open X-Embodiment — Robotic Learning Datasets and RT-X Models. arXiv:2310.08864.

[26] Temporal est un moteur d'orchestration de workflows open-source (Temporal Technologies, 2020) conçu pour maintenir l'état et l'historique complet des processus distribués. Il constitue un exemple paradigmatique des architectures d'event sourcing appliquées aux workflows longs. Sa mention ici est illustrative : la comparaison s'applique plus généralement à toute architecture fondée sur la conservation de l'historique événementiel complet d'un processus.

[24] Nigro, G., & Neisser, U. (1983). Point of view in personal memories. Cognitive Psychology, 15(4), 467–482.

[22] Cette gradation vise à situer les régimes informationnels par rapport à l'ancrage perceptif, non à établir une hiérarchie épistémologique générale. La relation entre théorisation scientifique et expérience empirique est elle-même complexe : la théorie structure le regard expérimental en retour (Kuhn, T., The Structure of Scientific Revolutions, 1962 ; Bachelard, G., La Formation de l'esprit scientifique, 1938). Le schéma proposé ici ne préjuge pas de cette relation — il positionne uniquement le régime informationnel des LLMs par rapport aux autres niveaux.

[19] Hafner, D. et al. (2023). Mastering Diverse Domains through World Models (DreamerV3). arXiv:2301.04104.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

[20] Friston, K. et al. (2022). Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press.

[21] Brohan, A. et al. (2023). RT-X: Open X-Embodiment — Robotic Learning Datasets and RT-X Models. arXiv:2310.08864.

[26] Temporal est un moteur d'orchestration de workflows open-source (Temporal Technologies, 2020) conçu pour maintenir l'état et l'historique complet des processus distribués. Il constitue un exemple paradigmatique des architectures d'event sourcing appliquées aux workflows longs. Sa mention ici est illustrative : la comparaison s'applique plus généralement à toute architecture fondée sur la conservation de l'historique événementiel complet d'un processus.

[24] Nigro, G., & Neisser, U. (1983). Point of view in personal memories. Cognitive Psychology, 15(4), 467–482.

[22] Cette gradation vise à situer les régimes informationnels par rapport à l'ancrage perceptif, non à établir une hiérarchie épistémologique générale. La relation entre théorisation scientifique et expérience empirique est elle-même complexe : la théorie structure le regard expérimental en retour (Kuhn, T., The Structure of Scientific Revolutions, 1962 ; Bachelard, G., La Formation de l'esprit scientifique, 1938). Le schéma proposé ici ne préjuge pas de cette relation — il positionne uniquement le régime informationnel des LLMs par rapport aux autres niveaux.

Pour aller plus loin : Eléments bibliographiques

Barsalou, L. W. (1999). Perceptual Symbol Systems. Behavioral and Brain Sciences, 22(4), 577–609.

Clark, A., & Chalmers, D. (1998). The Extended Mind. Analysis, 58(1), 7–19.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8(2), 240–247.

Cytowic, R. E., & Eagleman, D. M. (2009). Wednesday Is Indigo Blue. MIT Press.

Damasio, A. (1994). Descartes' Error: Emotion, Reason, and the Human Brain. Putnam.

Dreyfus, H. (1972). What Computers Can't Do. Harper & Row.

Husserl, E. (1939). Erfahrung und Urteil. Acad. Verlagsgesellschaft. Trad. fr. : Expérience et Jugement, PUF, 1970.

Dreyfus, H. (1991). Being-in-the-World. MIT Press.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D, 42(1–3), 335–346.

Held, R. et al. (2011). The newly sighted fail to match seen with felt. Nature Neuroscience, 14(5), 551–553.

Jackson, F. (1982). Epiphenomenal Qualia. The Philosophical Quarterly, 32(127), 127–136.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview.

Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML 2021.

Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–424.

Spelke, E. (1990). Principles of Object Perception. Cognitive Science, 14(1), 29–56.

Tonegawa, S. et al. (2015). Memory Engram Cells Have Come of Age. Neuron, 87(5), 918–931.

Nigro, G., & Neisser, U. (1983). Point of view in personal memories. Cognitive Psychology, 15(4), 467–482.

Rosch, E. (1973). Natural categories. Cognitive Psychology, 4(3), 328–350.

Temporal Technologies (2020). Temporal Workflow Engine. **https://temporal.io**. Cf. aussi l'Event Sourcing pattern : Fowler, M. (2005). Event Sourcing. **martinfowler.com**.

Tulving, E. (1983). Elements of Episodic Memory. Oxford University Press.

Tulving, E. (2002). Episodic Memory: From Mind to Brain. Annual Review of Psychology, 53, 1–25.

Varela, F., Thompson, E., & Rosch, E. (1991). The Embodied Mind. MIT Press.

Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS 2017.

Wittgenstein, L. (1953). Philosophical Investigations. Blackwell.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Bachelard, G. (1938). La Formation de l'esprit scientifique. Vrin.

Brohan, A. et al. (2023). RT-X: Open X-Embodiment. arXiv:2310.08864.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com

Friston, K. et al. (2022). Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press.

Hafner, D. et al. (2023). Mastering Diverse Domains through World Models (DreamerV3). arXiv:2301.04104.

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

Twingital-institute / Twingital-ventures : twingital-ventures.com