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

Jérôme Vetillard

Twingital Institute — Working Paper

2025

Résumé

La critique classique selon laquelle les systèmes d'intelligence artificielle seraient structurellement enfermés dans le langage repose sur une confusion conceptuelle entre langage et représentation. Le problème n'est pas d'abord que ces systèmes manipulent du langage, mais la manière dont leurs représentations sont formées, ancrées et organisées. Toute intelligence, qu'elle soit biologique ou artificielle, semble impliquer une médiation représentationnelle préalable à tout traitement de l'information. Cette thèse n'implique pas nécessairement l'existence de représentations symboliques explicites : les formes minimales de cognition peuvent reposer sur des médiations subsymboliques, distribuées ou dynamiques. Elle implique en revanche qu'aucun système cognitif ne traite le monde brut sans transformation préalable de ses signaux en états internes exploitables. L'article introduit alors une distinction que la littérature technique tend souvent à sous-estimer : dans les systèmes biologiques, l'encodage est intrinsèquement transductif, résultant d'une conversion directe de signaux physiques au sein de boucles continues perception–action ; dans les systèmes d'IA contemporains, les représentations sont le plus souvent apprises à partir de données préalablement collectées, sélectionnées et structurées par des humains. Même l'adjonction de capteurs physiques ne résoudrait que le premier niveau du grounding. Deux niveaux plus profonds demeureraient ouverts : un grounding multimodal fondé sur la co-constitution de plusieurs modalités perceptives au sein d'une même trajectoire d'action située, et un grounding épisodique et affectif fondé sur une organisation biographique de la mémoire — architecture mnésique distribuée, multimodale et biographiquement structurée, dont les liaisons sont forgées par co-expérience vécue et modulées par une valence affective. C'est cette architecture qui constitue aujourd'hui l'écart architectural le plus profond entre cognition humaine et systèmes d'intelligence artificielle contemporains — un écart non ontologique mais architectural et historique, dont les conditions de résorption restent à ce jour non instanciées.

Mots-clés : *encodage, transduction, grounding, mémoire épisodique, graphe mnésique, synesthésie, world model, cognition incarnée, LLM, engram distribué.*

1. L'objection classique et une reformulation plus rigoureuse

Une critique récurrente adressée aux systèmes d'intelligence artificielle consiste à affirmer qu'ils seraient structurellement limités par leur dépendance au langage. Selon cette perspective, l'IA ne ferait que manipuler des symboles ou des tokens, tandis que la cognition humaine serait

capable d'une forme de pensée pré-linguistique, parfois qualifiée d'*ante-prédicative*[2] — observable notamment chez le nourrisson avant l'acquisition du langage.

Cette objection s'inscrit dans une tradition philosophique large. Elle trouve des résonances dans le linguistic turn du XXe siècle et dans les analyses de Wittgenstein sur les jeux de langage — lesquels suggèrent que le sens émerge de pratiques sociales incarnées, non de représentations mentales privées. Elle s'articule également, de manière indirecte, à l'hypothèse Sapir-Whorf : si la langue que l'on parle contraint les catégories cognitives disponibles pour les humains, a fortiori un système dont les représentations sont intégralement construites à partir de corpus linguistiques serait-il enfermé dans les découpages conceptuels que ce corpus véhicule. Précisons toutefois que l'hypothèse Sapir-Whorf, dans ses formulations fortes, est aujourd'hui très discutée empiriquement — son usage ici vaut comme analogie structurelle, non comme fondement établi.

Hubert Dreyfus, dans *What Computers Can't Do*[3] (1972) puis dans *Being-in-the-World*[4] (1991), formule la critique la plus radicale de cette orientation. Inspiré de la phénoménologie heideggérienne, il soutient que la cognition humaine est fondamentalement incarnée et situationnelle : l'intelligence n'y apparaît pas comme la manipulation explicite de symboles abstraits, mais comme un *savoir-faire pratique* (*know-how*) émergeant de l'engagement direct dans un environnement vécu. La compréhension n'est pas, dans cette perspective, un calcul sur des représentations explicites ; elle est une manière d'être dans le monde. Aucune formalisation purement symbolique ne peut, selon Dreyfus, reproduire ce mode d'être.

La mante religieuse constitue un exemple minimal de ce type d'organisation. Elle capture des proies avec une efficacité remarquable sans posséder de représentation conceptuelle explicite de « proie », de « nutrition » ou de « métabolisme ». Son comportement repose sur des circuits sensorimoteurs spécialisés qui détectent certains motifs visuels et déclenchent une séquence motrice de capture. Un comportement adaptatif peut ainsi émerger sans concepts explicites ni modélisation propositionnelle du monde.

Cet exemple ne démontre cependant pas l'absence de toute médiation représentationnelle. Les circuits sensorimoteurs qui relient perception et action peuvent être interprétés, dans un cadre représentationnaliste élargi, comme des formes minimales de représentation : sub-symboliques, distribuées, dynamiques, inscrites dans l'organisation même de l'organisme. Cette reformulation ne recouvre pas nécessairement la position propre de Dreyfus, qui demeure réticent à l'usage du concept de représentation ; elle vise ici à traduire son intuition dans un vocabulaire plus compatible avec la discussion contemporaine des architectures cognitives.

Les LLM, en revanche, possèdent une compétence linguistique statistique très large mais restent dépourvus de ce type de couplage sensorimoteur direct. Ils manipulent des représentations dérivées du langage sans être engagés dans les boucles perception–action qui structurent l'apprentissage biologique.

À cette critique phénoménologique de l'action s'ajoute une critique d'ordre sémantique. John Searle, avec l'argument de la chambre chinoise[5], soutient qu'un système peut manipuler des symboles selon des règles purement syntaxiques tout en restant dépourvu de compréhension sémantique. La manipulation formelle de symboles ne garantit donc pas, en elle-même, l'accès

à la signification. Cet argument a suscité des réponses substantielles — notamment la *systems reply*, selon laquelle ce n'est pas le sous-système isolé mais le système complet qui comprendrait — sans que le débat soit aujourd'hui clos. Il identifie néanmoins une tension réelle que toute théorie de la cognition artificielle doit affronter : la distance entre traitement syntaxique et sémantique vécue.

Ces objections constituent un adversaire intellectuel sérieux. Elles identifient un problème réel dans les architectures actuelles. Toutefois, leur diagnostic peut être reformulé avec plus de précision.

Le problème fondamental n'est probablement pas le langage en tant que modalité cognitive. Les humains eux-mêmes mobilisent constamment des représentations symboliques et linguistiques dans leur cognition abstraite. La difficulté tient plutôt à l'absence d'ancrage causal entre les représentations internes d'un système et le monde physique — c'est-à-dire à l'absence de transduction : la capacité de transformer directement des signaux physiques issus de l'environnement, tels que la lumière, le son ou les variations proprioceptives, en structures représentationnelles internes.

Plus profondément encore, les architectures actuelles manquent de ce que nous proposons d'appeler une *structure biographique de la mémoire* — notion que nous développerons aux sections 5 et 6 à partir de Tulving (1983), Damasio (1994) et Barsalou (1999). La cognition humaine ne se développe pas comme l'apprentissage d'un modèle statique sur un corpus donné, mais comme l'accumulation progressive d'expériences situées dans le temps, organisées par une continuité autobiographique qui structure la formation des concepts, la hiérarchisation des connaissances et la stabilisation des catégories perceptives.

La limitation des LLM ne relève donc pas principalement d'un enfermement dans le langage, mais de l'absence de trois propriétés structurelles que cet article s'attache à décrire et à articuler. Premièrement, un ancrage sensorimoteur direct dans le monde physique, que nous désignerons comme grounding transductif. Deuxièmement, un grounding multimodal fondé sur la co-constitution vécue de plusieurs modalités perceptives au sein d'une même trajectoire d'action située, plutôt que sur leur simple co-occurrence statistique dans un corpus. Troisièmement, une mémoire épisodique organisée temporellement et modulée par des marqueurs affectifs, c'est-à-dire une architecture biographique. Le débat ne porte donc pas sur la possibilité abstraite d'une intelligence artificielle linguistique, mais sur la nécessité d'une architecture cognitive capable d'intégrer perception, action et mémoire biographique dans un même système dynamique. Cette question rejoint partiellement les travaux contemporains sur les world models, sans toutefois s'y réduire.

2. La représentation comme condition de toute cognition

Qu'elle soit biologique ou artificielle, toute forme d'intelligence semble impliquer une transformation préalable du monde en états internes manipulables. Un système cognitif n'agit pas directement sur le monde dans sa nudité physique ; il agit sur une structure informationnelle qui en sélectionne, condense et organise certains aspects pertinents.

Cette idée correspond à la position représentationnaliste classique en sciences cognitives, selon laquelle la cognition opère sur des représentations internes. Elle n'est toutefois pas universellement acceptée : certaines approches de la *cognition incarnée*, de l'énaction ou des systèmes dynamiques soutiennent que certains comportements intelligents peuvent émerger de boucles perception–action sans modèle interne explicite. La thèse défendue ici consiste moins à imposer une conception symboliste de la représentation qu'à affirmer qu'aucun traitement cognitif n'est possible sans transformation préalable des signaux du monde en états internes compatibles avec l'organisation du système. Nous reviendrons sur ce débat en §8.

Dans un cadre computationnel, cette transformation correspond à ce que l'on appelle l'encodage : la conversion d'une observation du monde en un objet mathématique exploitable, qu'il s'agisse d'un vecteur, d'une matrice ou d'un tenseur.

Une analogie fonctionnelle peut être établie avec la thèse kantienne des conditions de possibilité de l'expérience : de même que les formes *a priori* de la sensibilité organisent les données empiriques avant toute connaissance, les architectures computationnelles organisent leurs signaux d'entrée avant toute opération cognitive. Il ne s'agit évidemment pas d'identifier encodage computationnel et transcendantalisme kantien, mais de souligner une homologie fonctionnelle : dans les deux cas, l'accès au monde est médié par une structure organisatrice préalable.

Cette transformation peut intervenir à plusieurs niveaux de la chaîne de traitement : lors de la transduction sensorielle, dans les procédures de préparation des données supervisées par les ingénieurs, ou encore dans les couches internes d'un réseau de neurones qui apprennent leurs propres représentations latentes. La distinction entre ces niveaux n'est pas seulement technique ; elle est épistémologiquement décisive, car chacun introduit des hypothèses différentes sur la structure du monde que le système est susceptible de modéliser — et des acteurs différents qui exercent ce choix.

2.1 L'encodeur explicite : un module architectural dédié

Dans certaines architectures, l'encodeur constitue un module formellement identifiable et distinct du reste du modèle. L'exemple paradigmatique est l'architecture encodeur–décodeur, présente sous différentes formes dans de nombreux modèles neuronaux contemporains.

Dans les Transformers originaux (Vaswani et al., 2017), l'encodage commence par la transformation des tokens discrets en vecteurs continus au moyen d'une couche d'embedding. À ces vecteurs s'ajoute un encodage positionnel destiné à préserver l'information d'ordre dans la séquence — condition nécessaire au traitement de données séquentielles par une architecture par ailleurs permutation-invariante. Ces représentations initiales sont ensuite transformées par une pile de couches d'attention multi-têtes et de réseaux *feed-forward*. L'encodeur produit ainsi une séquence de représentations contextuelles dans un espace latent de grande dimension : chaque vecteur ne représente plus un token isolé, mais le token replacé dans l'ensemble de son contexte séquentiel. Le décodeur utilise ensuite ces vecteurs contextuels comme base informationnelle pour la génération.

Dans certains modèles tels que BERT (Devlin et al., 2018), seule la partie encodeur est conservée. Le modèle produit des représentations contextuelles destinées à être exploitées directement pour des tâches aval — classification, extraction d'information, recherche sémantique. L'espace latent produit par l'encodeur devient alors l'objet principal d'intérêt computationnel.

Les autoencodeurs constituent un autre exemple emblématique d'encodage explicite. Dans un autoencodeur variationnel (VAE, Kingma & Welling, 2013), l'encodeur projette les données d'entrée dans un espace latent régularisé probabilistiquement. Plutôt que de produire une représentation déterministe unique, le modèle apprend les paramètres d'une distribution latente — le plus souvent gaussienne — à partir de laquelle des échantillons peuvent être tirés. Cette contrainte agit comme un goulot d'étranglement informationnel qui force le modèle à capturer les structures les plus informatives des données. La régularisation probabiliste impose à cet espace une géométrie continue et exploitable : il devient possible d'interpoler entre deux représentations, d'explorer des directions latentes particulières ou de générer de nouvelles observations par échantillonnage. Cette navigabilité évoque, par analogie fonctionnelle, certaines propriétés de la mémoire humaine — cette analogie demeure toutefois heuristique : les mécanismes computationnels d'un espace latent génératif et l'organisation neurocognitive de la mémoire humaine relèvent de niveaux explicatifs distincts.

Les architectures multimodales contemporaines étendent ce principe en utilisant plusieurs encodeurs spécialisés traitant chaque modalité en parallèle. Des modèles comme CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022) ou Gemini reposent ainsi sur des encodeurs distincts pour le texte, l'image ou l'audio. L'entraînement vise à aligner ces représentations hétérogènes dans un espace latent commun : une image et sa description textuelle sont optimisées pour occuper des positions proches dans cet espace partagé. L'encodage devient ainsi simultanément modulaire — chaque modalité disposant de son propre encodeur spécialisé — et soumis à une contrainte d'alignement intermodal. Cette stratégie peut être interprétée comme une tentative technique de rapprocher des représentations symboliques et perceptuelles au sein d'un même espace informationnel, constituant une forme partielle de grounding.

Dans ces architectures, l'encodage ne constitue donc pas simplement une étape technique de transformation des données. Il définit la géométrie informationnelle dans laquelle le système peut comparer, raisonner et générer de nouvelles observations. La structure de cet espace latent conditionne directement ce que le modèle est capable d'apprendre, de généraliser et d'inférer.

2.2 L'encodage implicite : les couches d'un réseau comme encodeurs fonctionnels

Dans d'autres architectures, l'encodage n'apparaît pas comme un module distinct mais se trouve distribué dans les couches du réseau, qui apprennent progressivement des représentations de complexité croissante. Ce phénomène est particulièrement bien documenté dans les réseaux de neurones convolutionnels (CNN).

Dans un CNN profond, de type VGG, ResNet ou EfficientNet, les premières couches convolutionnelles apprennent généralement des filtres sensibles à des motifs visuels élémentaires tels que des orientations de contours, des gradients d'intensité ou certaines fréquences spatiales. Les couches intermédiaires combinent ces primitives pour former des motifs plus complexes — textures, structures répétitives, fragments d'objets. Les couches profondes produisent enfin des représentations corrélées à des catégories sémantiques plus globales. Cette organisation hiérarchique présente une analogie fonctionnelle avec la hiérarchie de traitement du cortex visuel ventral humain, souvent schématisée par la progression V1 → V2 → V4 → IT. Elle s'en distingue toutefois fondamentalement : dans les systèmes biologiques, ces transformations reposent sur des dynamiques neuronales complexes, des boucles récurrentes, des modulations attentionnelles et des contraintes d'apprentissage biologiquement déterminées. Dans les CNN, elles émergent d'un apprentissage par descente de gradient sous la seule contrainte de la fonction de perte — l'encodage est une propriété *émergente* du processus d'optimisation, non un module explicitement conçu.

Dans les réseaux récurrents de type LSTM ou GRU, chaque couche maintient un état caché qui constitue une représentation dynamique du contexte accumulé, mise à jour à chaque pas de séquence selon des pondérations apprises. L'encodage y est ainsi temporellement distribué plutôt que spatialement localisé.

Dans les Transformers, l'encodage émerge de l'alternance, à chaque couche, entre un mécanisme d'auto-attention et un réseau *feed-forward* non linéaire. L'auto-attention recalcule à chaque couche une représentation contextualisée de chaque token par agrégation pondérée des autres tokens de la séquence. Le réseau *feed-forward* intercalé opère ensuite une transformation non linéaire position par position, permettant au modèle de construire des représentations qui ne sont pas réductibles à de simples moyennes pondérées contextuelles. Dans les architectures encodeur seul comme BERT, l'encodage utilisé pour les tâches aval est souvent dérivé de la représentation finale du token spécial [CLS], après un traitement distribué sur l'ensemble des couches : l'encodage se concentre en un point après une transformation hiérarchique. Dans les architectures décodeur seul — telles que les modèles GPT et la plupart des LLMs contemporains — il n'existe pas de tel point de concentration : il n'y a pas d'encodeur séparé, et la représentation contextuelle est reconstruite à chaque génération de token depuis les couches du décodeur lui-même. L'encodage est ici intégralement distribué, recalculé dynamiquement à chaque étape de génération.

2.3 L'encodage dans le pipeline de données : la couche du data scientist

Une troisième forme d'encodage, souvent négligée dans les discussions théoriques sur l'intelligence artificielle, intervient en amont de tout modèle : l'encodage réalisé dans le pipeline de préparation des données, sous la supervision directe d'un data scientist ou d'un ingénieur de données. Cette couche est épistémologiquement la plus chargée, car elle détermine quelles dimensions du monde seront accessibles au modèle, sous quels formats et selon quelles hypothèses implicites.

Les décisions qui la composent ne sont pas toutes de même nature épistémique. Au premier niveau se situent les décisions portant sur le périmètre de la représentation : sélection des

variables, choix du schéma de données, définition des entités pertinentes, granularité temporelle ou spatiale des observations. C'est à ce niveau que se joue l'essentiel : le modèle ne peut pas identifier explicitement une variable absente de son espace de représentation. Pour lui, cette variable n'existe pas comme dimension possible du problème. Au deuxième niveau se situent les décisions portant sur la transformation des variables retenues : encodage des variables catégorielles (one-hot encoding, ordinal encoding, target encoding), vectorisation du texte (TF-IDF, embeddings lexicaux, représentations contextuelles), normalisation et standardisation des variables continues. Ces opérations définissent la géométrie mathématique dans laquelle le phénomène sera présenté au modèle. Au troisième niveau se situent les décisions portant sur l'incomplétude des données : imputation des valeurs manquantes, traitement des absences, exclusion ou agrégation de certaines observations. Ces opérations impliquent une théorie implicite de ce que signifie l'absence d'information.

Chacune de ces décisions est prise par des agents humains, sur la base de leur expertise, de leurs hypothèses sur la structure du problème et de leurs contraintes computationnelles. La pertinence de ces choix dépend directement de la connaissance du domaine étudié. En conséquence, le modèle entraîné sur ces données n'a aucun accès aux dimensions du monde qui n'ont pas été encodées. Cette opacité est structurelle : c'est au niveau de la préparation et de la structuration des données que se définit en grande partie l'espace de perception du modèle.

Cette situation diffère de la contrainte biologique. Dans les systèmes vivants, la physiologie sensorielle limite certes ce qui peut être transducté — un humain ne perçoit pas les ultraviolets, contrairement à l'abeille. Mais cette limitation résulte de contraintes phylogénétiques et physiologiques incorporées à l'organisme, non de décisions conceptuelles explicites prises par un agent extérieur au système avant son développement. Dans les systèmes artificiels, au contraire, l'espace de représentation est en grande partie déterminé par des choix humains qui précèdent l'apprentissage — nous y reviendrons en §3 — et que le modèle ne peut ni évaluer ni remettre en question.

Conclusion de section

Conclusion de la section 2

Dans tous les cas — encodage explicite, implicite ou pré-modèle — la logique fondamentale demeure identique : un système ne peut traiter l'information qu'après l'avoir convertie dans un format de représentation compatible avec ses opérations internes.

Dans les architectures d'intelligence artificielle contemporaines, ces représentations prennent généralement la forme d'espaces latents vectoriels. Une propriété importante de ces espaces est que les relations entre observations y deviennent calculables : des observations sémantiquement proches tendent à être projetées dans des régions voisines, tandis que des corrélations ou des structures hiérarchiques se traduisent sous forme de relations géométriques. Cette propriété est générale à tout espace latent appris — elle a été rendue célèbre par les modèles de type Word2Vec, où des opérations comme $roi - homme + femme \approx reine$ peuvent être réalisées directement dans l'espace vectoriel. Elle est particulièrement saillante dans les architectures multimodales comme CLIP, où des objets perceptifs et leurs descriptions

linguistiques sont projetés dans un espace latent partagé dont la proximité géométrique reflète la correspondance sémantique.

Cette géométrisation de l'information peut évoquer, à première vue, certaines propriétés des systèmes cognitifs biologiques, où des populations neuronales encodent également des relations structurelles entre stimuli. Toutefois, ce rapprochement doit être manié avec prudence : derrière l'apparente similarité des représentations se cachent des mécanismes de formation, des contraintes d'apprentissage et des formes d'ancrage au monde profondément différents. En particulier, les espaces de représentation des systèmes artificiels sont largement déterminés par les choix d'encodage qui précèdent ou structurent l'apprentissage — qu'ils soient architecturaux ou introduits dans le pipeline de données. C'est précisément cette différence d'ancrage et de formation des représentations que les sections suivantes se proposent d'examiner.

3. L'asymétrie fondamentale : transduction biologique et encodage médiatisé

3.1 La transduction biologique : un encodage contraint par la physique et l'évolution

Dans les systèmes biologiques, l'encodage des informations issues de l'environnement est assuré par des transducteurs sensoriels — des structures anatomiques spécialisées capables de convertir certaines formes d'énergie physique (photons, vibrations mécaniques, gradients chimiques, pression, accélérations) en signaux électrochimiques exploitables par le système nerveux.

L'oreille interne constitue un exemple paradigmatique. Les cellules ciliées de la cochlée convertissent des vibrations mécaniques en potentiels d'action par la déflexion des stéréocils et l'ouverture de canaux ioniques mécanosensibles. Les propriétés physiques du stimulus sonore — notamment sa fréquence — sont ainsi traduites en organisation spatiale de l'activité neuronale le long de la membrane basilaire selon le principe de tonotopie. À ces informations auditives s'ajoutent celles du système vestibulaire, qui encode les accélérations angulaires et linéaires ainsi que l'orientation gravitationnelle. Ces signaux sont intégrés avec les informations visuelles et proprioceptives afin de stabiliser la posture et coordonner l'action en temps réel[5].

Ces mécanismes ne doivent toutefois pas être compris comme un accès transparent au monde. Bien avant d'atteindre les aires corticales, le signal a déjà subi de multiples transformations : filtrage rétinien, inhibition latérale, normalisation, sélection, compression temporelle et spatiale. Les systèmes sensoriels biologiques effectuent donc eux aussi une transformation substantielle du signal. La différence essentielle ne réside pas dans l'existence d'un encodage — tout système cognitif encode — mais dans l'origine des contraintes qui structurent cet encodage. Dans les organismes vivants, ces contraintes résultent de régularités biophysiques, de pressions évolutives et d'interactions écologiques prolongées avec un environnement physique. Les transducteurs sensoriels ont été façonnés par l'histoire phylogénétique de l'espèce et par les propriétés statistiques des niches écologiques au sein desquelles elle s'est développée.

Dans ce cadre, perception et action forment des boucles dynamiques continues. La perception ne consiste pas seulement à produire une description interne du monde, mais à détecter des possibilités d'action pertinentes dans l'environnement. Dans le vocabulaire de James Gibson, les systèmes perceptifs sont couplés aux *affordances*[6] de l'environnement — les opportunités d'action qu'un environnement donné offre à un organisme donné. Il convient de préciser que Gibson défend une position explicitement anti-représentationnaliste : pour lui, les *affordances* sont directement perçues sans médiation représentationnelle. L'usage qui en est fait ici s'inscrit dans une lecture représentationnaliste minimale compatible avec l'argument général de cet article : les *affordances* sont comprises comme des propriétés relationnelles émergeant du couplage dynamique entre organisme et environnement. Ce type de couplage perception–action constitue précisément ce que les modèles de langage contemporains, dépourvus de boucle sensorimotrice fermée, ne peuvent instancier.

3.2 L'encodage artificiel : une chaîne de médiations représentationnelles

Dans les systèmes d'intelligence artificielle contemporains — et en particulier dans les modèles fondés sur des corpus symboliques tels que les grands modèles de langage — l'encodage des données obéit à une logique profondément différente. Avant que le modèle ne puisse traiter la moindre information, les données ont déjà traversé une chaîne de médiations humaines impliquant ingénieurs de données, data scientists et annotateurs, dont les principales opérations ont été décrites au §2.3. Ces transformations n'ont pas pour fonction de reproduire directement la structure physique du monde, mais de produire une représentation exploitable par une architecture d'apprentissage statistique. Le modèle n'est donc pas confronté aux propriétés physiques de l'environnement, mais à des artefacts informationnels déjà produits par l'activité cognitive humaine.

Il serait toutefois inexact d'affirmer que les systèmes d'IA sont dépourvus de contraintes. Ils sont au contraire fortement contraints par plusieurs facteurs : la structure de leur architecture, la forme de la fonction de perte qui oriente l'apprentissage, la distribution statistique des données d'entraînement, les biais inductifs introduits par l'architecture elle-même, ainsi que les capacités de calcul disponibles. Ces contraintes jouent un rôle déterminant dans la formation des représentations internes. Cependant, elles diffèrent en nature des contraintes biologiques. Les contraintes biologiques sont le produit de processus évolutifs et d'interactions écologiques prolongées avec un monde physique ; les contraintes des systèmes artificiels sont définies par des ingénieurs et optimisées pour des objectifs de performance mesurables sur des ensembles de données donnés, le plus souvent sans interaction directe, continue et ouverte avec l'environnement auquel ces représentations se rapportent.

Dans le cas des modèles de langage, la médiation est particulièrement marquée. Les données d'entraînement sont constituées de textes qui représentent déjà des descriptions, interprétations et conceptualisations humaines du monde. La chaîne de transformation peut être schématisée de la manière suivante :

Biologique : monde physique → transduction sensorielle → signal neural → cognition

```
LLM :        monde physique → expérience humaine → description
symbolique → corpus → tokenisation → modèle
```

Le système d'IA opère ainsi sur des représentations déjà stabilisées dans des artefacts sémiotiques. Le signal traité par le modèle n'est pas un flux énergétique issu directement de l'environnement, mais un objet symbolique produit par des agents humains. En ce sens, on peut parler d'un *encodage de second degré* : dans les systèmes biologiques, l'encodage correspond à une transformation de signaux physiques issus de l'environnement ; dans les systèmes d'IA fondés sur des corpus symboliques, l'apprentissage porte sur des représentations humaines du monde déjà constituées. Le modèle apprend à modéliser les structures discursives par lesquelles les humains décrivent, interprètent et organisent le monde, plutôt que les propriétés physiques du monde lui-même.

Cette distinction ne doit cependant pas être absolutisée. La cognition humaine adulte elle-même est traversée de médiations culturelles, linguistiques, sociales et institutionnelles. Un individu humain n'accède pas au monde uniquement par perception directe : il hérite aussi d'un monde déjà structuré par des récits, des catégories, des instruments, des pratiques et des institutions. La différence ici défendue ne réside donc pas dans l'opposition simpliste entre une cognition biologique supposément immédiate et une cognition artificielle entièrement médiée, mais dans la nature et dans l'empilement des régimes de médiation qui président à la formation des représentations.

3.3 Distance épistémique au monde

Cette différence peut être formulée plus généralement en termes de *distance épistémique au monde*. Par distance épistémique, on entend ici le nombre et la nature des médiations informationnelles séparant un état interne du système des propriétés physiques de l'environnement. Dans les systèmes biologiques, cette chaîne est relativement courte : certaines propriétés énergétiques du monde sont directement transduites en activité neuronale, puis intégrées dans des boucles perception–action contraintes par la physiologie et l'évolution. Dans les systèmes d'IA corpus-based, cette chaîne est à la fois plus longue et qualitativement différente : le signal a déjà été collecté, sélectionné, filtré, structuré et représenté par des agents humains avant d'être introduit dans le système d'apprentissage. Surtout, la nature du signal change en cours de route — il passe d'un régime physique à un régime symbolique.

> *Les systèmes d'IA dominants, lorsqu'ils sont entraînés principalement sur des corpus textuels, apprennent donc avant tout à partir de représentations symboliques du monde, et non à partir d'une interaction directe avec celui-ci.*

Cette distance épistémique accrue ne constitue pas nécessairement une limitation absolue : elle permet au contraire d'exploiter l'immense accumulation de connaissances symboliques produites par les sociétés humaines, ce qui explique l'efficacité remarquable de ces modèles sur de nombreuses tâches propositionnelles. Mais elle implique que les représentations internes des systèmes artificiels sont formées dans un régime informationnel profondément différent de celui des organismes biologiques. Cette asymétrie constitue le premier niveau de la différence architecturale que cet article cherche à cartographier. Comme on le verra dans les sections suivantes, deux niveaux supplémentaires demeurent ouverts même pour un système artificiel

équipé de transducteurs physiques : un grounding multimodal fondé sur la coordination vécue des modalités perceptives, et un grounding autobiographique et affectif reposant sur une organisation biographique de la mémoire.

Tableau 1 — Comparaison des modes d'encodage biologique et artificiel

Dimension	Encodage biologique (transduction)	Encodage artificiel (LLMs)
Source de données	Expérience sensorimotrice directe	Texte produit et filtré par des humains
Mode d'encodage	Transductif — ancré dans la physique	Symbolique/statistique, médiatisé
Représentation de la causalité	Modèle causal structurel	Corrélation statistique
Temporalité	Dynamique continue, incarnée	Séquence de tokens
Grounding	Ancré dans l'action et la perception	Absent — symboles autoreférentiels
Incarnation	Constitutive	Absente des architectures actuelles

Sources : Gibson (1979), Harnad (1990), LeCun (2022) — synthèse de l'auteur.

4. Le symbol grounding problem et les formes d'aveuglement représentationnel

4.1 Le symbol grounding problem

L'asymétrie décrite dans la section précédente rejoint directement le *symbol grounding problem* formulé par Stevan Harnad en 1990[7]. Dans un système purement formel, les symboles ne renvoient initialement qu'à d'autres symboles : leur signification est définie par des relations internes au système et non par un ancrage direct dans le monde physique. Un dictionnaire fournit une illustration simple de ce phénomène : chaque mot y est défini par d'autres mots, et la signification globale ne peut émerger que si certains symboles sont finalement reliés à une expérience non symbolique. Cette situation est une conséquence directe de la distance épistémique introduite à la section précédente : le système apprend non pas à partir d'interactions physiques avec l'environnement, mais à partir de représentations symboliques produites par des agents humains ayant eux-mêmes expérimenté le monde.

Les grands modèles de langage se trouvent précisément dans cette situation. Ils apprennent des structures linguistiques extrêmement riches à partir de corpus textuels massifs. Mais ces symboles renvoient principalement à d'autres symboles produits par l'activité cognitive humaine. Le modèle construit ainsi un espace relationnel dense entre concepts sans disposer, en régime purement textuel, d'un accès direct aux phénomènes auxquels ces concepts se rapportent. Sa représentation du monde est dérivée de descriptions linguistiques produites par des agents humains qui, eux, ont interagi avec le monde.

Il convient de préciser que, chez Harnad, le grounding ne se réduit pas à une simple association entre un symbole et une mesure physique brute. Ce qui est en jeu est l'ancrage des symboles dans des catégories perceptuelles apprises, issues d'interactions sensorimotrices permettant de

discriminer de manière robuste des entités ou des propriétés du monde. Le problème n'est donc pas seulement l'absence de contact avec le réel, mais l'absence d'un processus par lequel les symboles seraient reliés à des catégories acquises sur la base d'une expérience perceptive.

4.2 Aveuglement perceptif : le problème de Molyneux et l'argument de Mary

La distinction entre connaissance conceptuelle et expérience perceptive apparaît clairement dans deux traditions philosophiques distinctes.

Le *problème de Molyneux*, formulé en 1688, demandait si un aveugle de naissance capable de distinguer par le toucher un cube d'une sphère pourrait reconnaître ces objets par la vision s'il recouvrait soudainement la vue. Les recherches empiriques menées par Richard Held et ses collègues, publiées en 2011 dans *Nature Neuroscience*[8], ont examiné des patients ayant recouvré la vue après une cataracte congénitale. Les résultats indiquent que les sujets ne peuvent pas immédiatement identifier visuellement des formes qu'ils reconnaissaient auparavant par le toucher. La correspondance entre les modalités perceptives doit être progressivement apprise par l'expérience. Ces travaux suggèrent que la connaissance conceptuelle d'un objet et l'expérience perceptive de cet objet constituent deux types de représentations fonctionnellement distincts.

L'argument de Mary, proposé par Frank Jackson en 1982[9], radicalise cette distinction. Mary est une scientifique spécialiste de la neurophysiologie de la vision des couleurs qui a vécu toute sa vie dans un environnement monochromatique. Elle connaît toutes les lois physiques et neurobiologiques impliquées dans la perception du rouge. Pourtant, lorsqu'elle voit une surface rouge pour la première fois, il semble qu'elle acquière une forme de connaissance nouvelle : la connaissance phénoménologique de ce que signifie voir du rouge — ce que Jackson nomme le *qualia* correspondant. Cet argument demeure débattu dans la philosophie contemporaine de l'esprit — les objections fonctionnalistes de Lewis, Dennett et Levin en contestent les prémisses — mais il met en évidence une distinction qui reste pertinente pour la discussion présente : celle qui sépare la connaissance propositionnelle d'un phénomène de son expérience perceptive.

Une objection classique consiste à remarquer qu'un individu peut comprendre la douleur d'autrui sans l'éprouver lui-même. Cette compréhension relève toutefois d'une inférence empathique ancrée dans des douleurs déjà vécues et dans l'observation de réactions humaines : elle ne donne pas accès à la qualité phénoménologique précise de cette douleur ni à son intensité vécue, mais produit une modélisation conceptuelle plausible du phénomène.

Appliqué aux modèles de langage, le parallèle devient clair. Un LLM entraîné sur des milliards de descriptions textuelles du rouge peut en décrire la physique, la neurophysiologie et les associations culturelles. Il peut décrire la douleur et ses effets comportementaux. Mais, en régime purement textuel, il ne dispose d'aucune expérience perceptive correspondante. Sa représentation demeure dérivée de descriptions linguistiques produites par des agents humains.

4.3 La théorisation scientifique ne s'affranchit pas de l'ancrage empirique

On pourrait toutefois objecter que l'histoire des sciences montre que des structures du monde peuvent être inférées sans perception directe. Au Ve siècle avant notre ère, Démocrite propose que la matière soit constituée d'unités indivisibles alors qu'aucun instrument ne permet d'observer des structures microscopiques. De même, la théorie de la relativité restreinte formulée par Einstein en 1905 mobilise largement des expériences de pensée — la poursuite d'un rayon lumineux ou la comparaison d'horloges en mouvement — avant que certaines de ses prédictions ne soient confirmées expérimentalement. Ces exemples pourraient suggérer que la connaissance scientifique peut émerger indépendamment de toute perception directe.

Une analyse plus attentive montre cependant que les expériences de pensée ne remplacent pas l'expérience empirique : elles réorganisent conceptuellement des régularités déjà observées. L'atomisme antique s'inscrit dans une réflexion sur la divisibilité des substances et sur des phénomènes macroscopiques observables. La relativité restreinte se développe dans un contexte structuré par l'électromagnétisme de Maxwell et par les mesures de Michelson et Morley (1887). Les théories scientifiques émergent au sein de ce que l'on peut appeler un état socio-technique des sciences — un environnement constitué d'instruments, de méthodes expérimentales, de résultats accumulés et de cadres conceptuels partagés. La créativité scientifique consiste moins à produire des structures symboliques indépendamment du monde qu'à inférer des structures invisibles à partir de régularités observables.

4.4 Gradation épistémique

Cette analyse permet de situer les modèles de langage sur une échelle de médiation par rapport à l'expérience physique[22]. Un physicien tel qu'Einstein infère des structures théoriques à partir d'un rapport indirect mais réel à l'expérience empirique, aux instruments et aux résultats produits par sa communauté. Un étudiant qui lit ses travaux sans jamais réaliser d'expérience possède déjà un accès plus médié à ces phénomènes : il hérite d'un savoir théorique déjà stabilisé par l'activité scientifique. Un modèle de langage se situe à un niveau supplémentaire de médiation : il n'interagit ni avec le monde physique ni avec les pratiques expérimentales de la science ; il apprend à modéliser les structures discursives produites par les humains qui décrivent ces expériences.

interaction empirique → élaboration théorique → description symbolique → modélisation statistique des textes

Il ne s'agit pas de soutenir que toute médiation éloigne nécessairement de la vérité. Le point est plus précis : à mesure que se multiplient les médiations, le système dépend davantage des formes symboliques déjà construites par d'autres agents pour organiser son rapport au monde. Les modèles de langage héritent ainsi d'un ancrage empirique humain, mais sous une forme seconde, sémiotisée et déjà interprétée.

4.5 Conséquence pour le grounding

La science humaine peut formuler des hypothèses sur des phénomènes invisibles précisément parce qu'elle repose, directement ou indirectement, sur un ancrage empirique préalable dans le monde physique. Les systèmes d'IA fondés sur des corpus symboliques héritent de cet ancrage de manière indirecte, à travers les traces discursives produites par les communautés humaines.

Ils peuvent modéliser avec une grande efficacité les relations entre ces représentations. Mais tant que certains symboles ne sont pas reliés à des interactions effectives avec l'environnement, leur signification demeure, au moins en première analyse, interne au réseau symbolique qui les relie. C'est précisément cette absence d'ancrage perceptif direct que met en évidence le symbol grounding problem, et que les différentes formes de grounding examinées dans les sections suivantes permettent de stratifier.

5. Les niveaux du grounding : du capteur à la mémoire autobiographique

5.1 Ce que résoudrait un capteur — et ce qu'il ne résoudrait pas

Une objection classique consiste à considérer que la limitation identifiée dans les modèles de langage n'est pas structurelle mais simplement architecturale : il suffirait d'équiper ces systèmes de capteurs physiques pour résoudre le problème du grounding. Un capteur optique capable de mesurer la distribution spectrale de la lumière incidente permettrait par exemple d'associer une mesure autour de 700 nm à la catégorie « rouge ». Le symbole linguistique serait alors relié à une propriété physique mesurable du monde. Cette intuition correspond précisément à celle qui sous-tend le symbol grounding problem formulé par Harnad : éviter que les symboles ne renvoient qu'à d'autres symboles en les ancrant dans des catégories perceptuelles apprises à partir d'interactions sensorimotrices avec l'environnement. Dans cette perspective, l'intégration de capteurs constitue effectivement une condition nécessaire pour dépasser la circularité purement symbolique.

Cependant, l'introduction d'un capteur ne résout qu'une partie du problème. Ce que fournit un capteur est une information sensorielle mesurable, à partir de laquelle un système peut apprendre à discriminer des catégories perceptuelles. Ce qu'il ne fournit pas à lui seul, c'est l'ensemble des structures cognitives associatives qui constituent un concept dans la cognition humaine.

Pour un être humain, la catégorie « rouge » ne correspond pas simplement à une propriété spectrale — et cette propriété spectrale n'est d'ailleurs, en pratique, pas explicitement représentée comme telle dans l'expérience ordinaire. Le rouge est intégré à un vaste réseau d'associations perceptives, motrices, culturelles, sociales et affectives : la texture et l'odeur d'une tomate coupée, le goût d'une sauce bolognaise, le souvenir d'un repas, la signification d'un feu rouge dans la circulation, la valence d'une alerte médicale, la charge symbolique de certaines couleurs dans un contexte culturel donné. Ces associations ne résultent pas uniquement de co-occurrences linguistiques ; elles sont stabilisées, enrichies et parfois réorganisées par la co-activation de multiples modalités perceptives et contextuelles au cours de l'expérience vécue.

Il serait toutefois excessif d'opposer ici de manière absolue apprentissage linguistique et expérience incarnée. Chez l'humain, les concepts se forment à l'intersection de plusieurs sources : expérience perceptive, action, langage, transmission sociale, culture et mémoire. Le point défendu ici est plus restreint : l'ajout de capteurs peut ancrer certains symboles dans des

propriétés du monde physique, mais il ne reproduit pas automatiquement l'ensemble des mécanismes par lesquels les représentations humaines acquièrent leur richesse multimodale, contextuelle et biographique.

5.2 Une stratification des niveaux de grounding

La littérature sur le grounding est souvent présentée comme si elle désignait un problème unique. L'analyse précédente suggère au contraire qu'il est utile de distinguer plusieurs niveaux conceptuellement distincts, chacun correspondant à une forme particulière de relation entre représentation et expérience.

Le premier niveau correspond au *grounding référentiel ou perceptuel* au sens de Harnad : relier les symboles à des catégories perceptuelles apprises à partir de données sensorimotrices. Les capteurs constituent ici un élément nécessaire, car ils permettent d'ancrer les représentations dans des propriétés mesurables de l'environnement. Ce niveau est partiellement soluble par l'adjonction de capteurs physiques, à condition que ces capteurs soient couplés à un mécanisme d'apprentissage catégoriel.

Le deuxième niveau correspond au *grounding multimodal*. Les concepts humains ne reposent pas sur une modalité sensorielle isolée mais sur l'intégration de multiples modalités. Barsalou (1999)[10] propose la théorie des *perceptual symbol systems*, selon laquelle les concepts sont représentés sous forme de simulateurs multimodaux capables de réactiver partiellement les circuits perceptifs et moteurs mobilisés lors de l'expérience initiale. Un concept comme « tomate » mobilise ainsi simultanément des représentations visuelles, olfactives, gustatives, tactiles et motrices. Ce qui forge ces liaisons n'est pas seulement la co-occurrence des mots dans un corpus, mais la coordination répétée de modalités hétérogènes dans des épisodes d'expérience situés. Ce niveau n'est pas dissous par l'ajout de capteurs pris isolément ; il suppose une histoire d'interactions multimodales suffisamment riches, synchronisées et contextualisées pour stabiliser ce type d'intégration.

Les architectures multimodales contemporaines commencent à explorer une partie de cet espace. Des systèmes vision-langage ou audio-vision peuvent apprendre certaines correspondances cross-modales à partir de données synchronisées — ce qui montre qu'une part du problème est, au moins fonctionnellement, abordable. Mais cette possibilité ne suffit pas à effacer l'écart avec l'intégration multimodale humaine, qui s'inscrit dans une boucle perception–action continue, dans un corps situé et dans une histoire d'apprentissage ouverte.

Le troisième niveau correspond au *grounding épisodique, affectif et autonoétique*. Il s'agit d'ancrer les représentations dans une mémoire personnelle structurée par des épisodes vécus. Endel Tulving (1983)[12] distingue trois composantes : la mémoire sémantique (*savoir que* — connaissances générales décontextualisées), la mémoire épisodique (*se souvenir de* — représentation mentale d'événements situés dans le temps et l'espace), et l'*autonoéticité* — la conscience de soi comme sujet ayant vécu cet événement, capable de se projeter mentalement dans le passé ou dans le futur. C'est ce troisième élément qui distingue le niveau 3 du niveau 2 : non seulement un réseau d'associations multimodales, mais un ancrage biographique dans lequel l'expérience est vécue depuis un point de vue subjectif situé. Cette dimension

autonoétique est renforcée par les marqueurs somatiques décrits par Damasio[13] : des signaux physiologiques attachés aux situations passées qui orientent implicitement les décisions et les associations futures. La tache sur la chemise n'est pas seulement une information perceptuelle — c'est une expérience datée, localisée, chargée d'affect et de subjectivité, qui a une place dans une vie. Ce niveau est structurellement inaccessible à tout système sans continuité biographique et sans autonoéticité.

Ces distinctions suggèrent que le grounding ne correspond pas à une propriété unique, mais à une hiérarchie de mécanismes cognitifs. L'ajout de capteurs peut résoudre une partie du problème en ancrant certaines représentations dans des propriétés physiques du monde. Toutefois, la richesse des concepts humains repose également sur l'intégration multimodale, sur l'organisation autobiographique de la mémoire et sur des dimensions autonoétiques et affectives que les architectures actuelles ne reproduisent pas de manière démontrée. Ce constat ne démontre pas une impossibilité de principe pour toute intelligence artificielle future ; il cartographie plutôt une série de seuils architecturaux croissants que l'équipement sensoriel, à lui seul, ne suffit pas à franchir.

Tableau 2 — Stratification des niveaux de grounding

Niveau	Nature	Contribution des capteurs	Références
1 — Grounding référentiel / perceptuel	Relier les catégories à des données sensorimotrices mesurables	Capteurs nécessaires mais non suffisants	Harnad (1990)
2 — Grounding multimodal	Intégration de multiples modalités via co-expérience incarnée	Capteurs utiles, nécessitent une expérience multimodale riche et située	Barsalou (1999)
3 — Grounding épisodique, affectif et autonoétique	Ancrage dans une mémoire autobiographique, affective et à perspective subjective	Non résolu par capteurs seuls ; requiert architecture de mémoire et autonoéticité	Tulving (1983), Damasio (1994)

Sources : Harnad (1990), Barsalou (1999), Tulving (1983), Damasio (1994) — synthèse de l'auteur.

Tableau 3 — Ce que le capteur résout et ne résout pas

Niveau de représentation	LLM textuel	Système multimodal avec capteurs	Cognition humaine
Associations linguistiques (rouge → sang)	✓ statistiques	✓ statistiques + perceptuelles	✓
Ancrage perceptuel direct	✗	✓ partiel	✓
Associations multimodales (rouge → goût bolognaise)	✗	✓ dépend données expérience	✓ via co-expérience incarnée
Mémoire d'événements individuels	✗	✓ traçabilité des interactions	✓ autobiographique et autonoétique
Marqueurs affectifs (Damasio)	✗	✗ (architectures actuelles)	✓

Autonoéticité (Tulving)	✗	✗	✓

Lecture : ✓ = présent ; ✗ = absent ou non démontré. Synthèse de l'auteur.

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 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 — 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 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.

Note phénoménologique de l'auteur

Ce que je décris ici n'est pas une abstraction théorique mais une donnée d'expérience directe. Je tisse mes souvenirs — et je peux me promener dans une trame qui est, pour moi, riche, dense et multidimensionnelle. Certains nœuds du graphe sont des arrêts sur image : des moments d'une densité sensorielle particulière, vibrants de plusieurs modalités simultanées, où la lumière d'une pièce, une odeur, une sensation musculaire, une émotion et une conversation se sont noués en un seul instant de présence. Ces nœuds ne sont pas des souvenirs parmi d'autres : ce sont des points d'ancrage à haute densité depuis lesquels le reste du tissu reste accessible. Je peux en outre changer de point de vue sur un même souvenir — le revisiter depuis l'entrée olfactive, puis visuelle, puis proprioceptive, puis affective — comme on tourne autour d'un objet dans l'espace pour en percevoir la profondeur. Cette mobilité de perspective dans l'espace mémoriel est, à ma connaissance, ce qui distingue le plus radicalement l'expérience subjective du souvenir vivant de toute forme actuelle de représentation computationnelle.

Statut épistémique de cette note : il s'agit d'une description phénoménologique à la première personne au sens husserlien — un donné d'expérience que la théorie doit expliquer, non une preuve de la thèse générale. Une expérience subjective, aussi précisément décrite soit-elle, ne démontre pas l'universalité du phénomène qu'elle illustre. Elle constitue ce que Merleau-Ponty (Phénoménologie de la perception, 1945) appelle un 'cas révélateur' : un point d'ancrage phénoménologique qui oriente la question théorique sans la résoudre. La question posée à la théorie est donc : quelles conditions architecturales permettraient à un système de produire et de naviguer dans une trame mémorielle de cette nature ?

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 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 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 majeur : 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*[16] (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 ambition n'est plus de modéliser seulement des relations symboliques, mais des régularités causales sous-jacentes aux transformations observables. Les world models constituent ainsi une tentative sérieuse de réduction de la distance épistémique au monde décrite dans les sections précédentes — ils répondent au moins partiellement au premier niveau du problème : celui de l'ancrage transductif des représentations.

7.2 Pourquoi cette avancée ne suffit pas

Cette avancée ne doit toutefois pas être surestimée. Les architectures multimodales contemporaines — CLIP, Gemini, Flamingo — réduisent partiellement l'écart entre modalités, mais leurs données demeurent sélectionnées, organisées et cadrées par des humains. Même lorsque des capteurs technologiques sont mobilisés, le pipeline de capture et de structuration reste largement médiatisé.

Les systèmes robotiques incarnés vont plus loin : ils interagissent physiquement avec des objets, reçoivent des retours haptiques, visuels et parfois proprioceptifs, et apprennent à partir de transitions causales produites par leurs propres actions. Des architectures comme RT-X[21], GR00T ou DreamerV3[19] représentent donc un changement réel par rapport aux LLM purement corpus-based. Mais ce changement, aussi important soit-il, ne suffit pas encore à faire émerger ce que ce texte nomme une architecture biographique. Ces systèmes peuvent apprendre des dynamiques, accumuler des épisodes, optimiser des politiques d'action et même maintenir certains états internes persistants. Ils ne montrent pas pour autant, de manière convaincante, la co-présence d'une continuité d'agent stable dans le temps, d'une mémoire épisodique persistante structurée autour d'un sujet, d'une modulation affective durable des liaisons mnésiques, et d'une récupération multi-perspective d'un même épisode. Autrement dit, ils approchent la transduction, parfois la multimodalité, mais pas encore la biographie.

7.3 Le seuil véritable : de l'état latent à l'arête biographique

C'est ici qu'apparaît le point central du présent article. Le problème n'est pas seulement de construire un système capable de prédire le monde. Il est de construire un système pour lequel

certaines relations internes portent la marque d'une co-expérience vécue par un agent continu. Dans un world model, les transitions entre états latents codent des régularités causales apprises à partir de l'expérience. Dans la mémoire humaine, les relations entre éléments mnésiques ne codent pas seulement des régularités causales ou statistiques : elles codent le fait que plusieurs dimensions d'expérience ont été vécues ensemble par un même sujet, dans un même contexte temporel, corporel et affectif. C'est cette différence que nous avons appelée l'arête biographique.

L'enjeu n'est donc pas de nier l'intérêt des world models. Il est de préciser leur statut : ils constituent très probablement la première étape nécessaire vers une cognition artificielle plus profondément ancrée, mais ils ne suffisent pas encore à reproduire la structure autobiographique de la mémoire humaine. Ils répondent au premier niveau de l'écart ; ils laissent largement ouverts les deux suivants.

7.4 Thèse reformulée

On peut désormais reformuler la thèse avec davantage de précision. L'écart entre cognition humaine et architectures contemporaines n'est pas ontologique. Il n'existe pas de raison de principe pour qu'un système artificiel ne puisse jamais instancier des structures de type biographique. Mais cet écart n'est pas non plus réductible à l'absence de capteurs ou à la pauvreté des corpus. Il est architectural : les systèmes actuels n'intègrent pas simultanément transduction, continuité d'agent, mémoire épisodique autonoétique, modulation affective et navigabilité multi-perspective.

8. Implications épistémologiques : ce que les systèmes actuels modélisent réellement

8.1 Ce que les LLM représentent effectivement

Les grands modèles de langage disposent d'une compétence propositionnelle étendue. Leurs espaces latents capturent des relations robustes présentes dans les corpus, permettant des analogies, des inférences et des généralisations d'une puissance remarquable. Il convient de reconnaître la réalité de cette compétence : ces systèmes modélisent effectivement des structures relationnelles riches, et ces structures permettent des usages cognitifs non triviaux.

Mais ce qu'ils 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* — Varela, Thompson & Rosch dans *The Embodied Mind*[17] (1991) ; Clark & Chalmers dans *The Extended Mind*[18] (1998) — ont souligné que l'intelligence émerge de l'interaction dynamique entre un organisme et son environnement, et que la cognition ne se réduit pas à une manipulation abstraite de symboles. 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. Il peut percevoir sans se souvenir comme sujet ; agir sans biographiser ses interactions ; accumuler des transitions sans les intégrer dans une histoire vécue. 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 du présent article est que la recherche en IA ne devrait pas seulement viser des systèmes plus performants, plus multimodaux ou plus autonomes. Elle devrait viser des systèmes dont les relations internes ne soient plus seulement statistiques ou logiques, mais historiques au sens vécu du terme. La question n'est pas simplement de savoir si un agent peut percevoir, prédire ou agir. Elle est de savoir s'il peut accumuler une histoire de ses interactions telle que certaines liaisons internes portent durablement la marque de cette histoire. Autrement dit, l'enjeu n'est pas seulement le world model. L'enjeu est le world model biographisé — si tant est que cette notion soit un jour opérationnalisable, ce qui reste une question ouverte.

9. Conclusion

L'argument développé dans cet article peut être reformulé de manière resserrée en quatre propositions.

Premièrement, toute forme de cognition suppose une médiation représentationnelle. L'objection classique selon laquelle l'IA serait prisonnière du langage repose sur une confusion : ce n'est pas le recours à des représentations qui constitue le problème, mais la manière dont ces représentations sont formées, ancrées et reliées.

Deuxièmement, l'encodage biologique et l'encodage des systèmes d'IA dominants diffèrent profondément par leur mode de couplage au monde. Dans les organismes vivants, l'encodage est transductif, intégré à des boucles perception–action façonnées par la biophysique et l'évolution. Dans les LLM, l'apprentissage repose principalement sur des corpus symboliques produits par des agents humains. Le système apprend donc à partir de représentations déjà médiatisées.

Troisièmement, le problème du grounding est stratifié. L'adjonction de capteurs ou la construction d'un world model permettent d'aborder le premier niveau, celui de l'ancrage transductif. Mais deux niveaux supplémentaires demeurent : un grounding multimodal fondé

sur la co-constitution vécue des modalités perceptives, et un grounding épisodique et affectif fondé sur une mémoire autobiographique.

Quatrièmement, la différence la plus profonde ne réside pas dans la taille des modèles, ni dans le nombre de modalités, ni même dans la présence de capteurs, mais dans la nature des liaisons internes entre représentations. La mémoire humaine repose sur des arêtes biographiques : des liaisons forgées par la co-expérience vécue d'un sujet situé, modulées par le contexte corporel, temporel et affectif, et récupérables depuis plusieurs perspectives. Aucune architecture actuelle — fût-elle multimodale, robotique ou fondée sur un world model — ne montre de manière convaincante la co-présence simultanée des conditions nécessaires à une telle architecture biographique : continuité d'agent, mémoire épisodique autonoétique persistante, modulation affective des liaisons, récupération multi-perspective.

L'écart n'est donc pas ontologique. Il n'y a pas de raison de principe pour qu'un système artificiel ne puisse jamais approcher une telle structure. Mais cet écart est aujourd'hui architectural et historique. C'est en ce sens précis que la question centrale posée par l'IA n'est pas une question de langage, ni même seulement une question de perception. C'est une question d'histoire : 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 concevoir un système qui n'ait pas seulement des données à traiter ou des états à enregistrer, mais quelque chose à se rappeler parce qu'il l'a vécu ?

Notes

[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.

[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.

[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.

[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.

Références 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.

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.

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.