VERS UNE INTELLIGENCE ARTIFICIELLE GENERALE : ÉPISTEMOLOGIE DE LA COGNITION SYNTHETIQUE ET INTRICATION COGNITIVE POUR L'HYBRIDATION NEURO-SYMBOLIQUE

Michael Pichat, Neocognition (Chrysippe R&D), Université de Paris & Facultés Libres de Philosophie et de Psychologie de Paris
Paloma Pichat, Neocognition (Chrysippe R&D)
Armanouche Gasparian, Neocognition (Chrysippe R&D)
contact@neocognition.ai

L'intelligence artificielle neuro-symbolique apparaît comme une prometteuse voie vers l'élaboration d'une nouvelle génération, beaucoup plus aboutie, d'intelligence artificielles générales.
Mais elle repose sur l'élaboration de systèmes originaux d'intrication fine des contenus et des processus cognitifs respectivement artificiels et humains, dont il s'agit de trouver des lieux (*topoï*) et des modes opératoires (*modi operandi*) de convergence. Dans une logique d'efficience, un tel couplage nécessite de nous clarifier plus avant sur la « nature » épistémologique d'un interfaçage « intime », à fine granularité, entre connaissances synthétiques et symboliques, afin de paramétrer de façon adaptée leur interaction ; et notamment, de mieux comprendre la dynamique cognitive propre qui caractérise la cognition synthétique.

1 Introduction

Deux grands types de dispositifs d'intelligence artificielle ont été historiquement développés à ce jour. La première est l'IA dite symbolique. D'un point de vue épistémologique, elle repose sur une conception cartésienne de l'intelligence synthétique, où cette dernière est façonnée à l'image de nos conceptions humaines de l'intelligence. Son principe est l'encodage synthétique de contenus et processus de rationalité : axiomes, raisonnements formels, éléments de logique, modélisations mathématiques, paramètres et contraintes, règles, lois physiques, ontologies, etc. Elle prend ainsi la forme d'un système dont l'être humain comprend et maîtrise directement le fonctionnement interne.

La deuxième est l'IA dite empirique. D'un point de vue épistémologique, elle s'apparente à une approche empiriste (humienne) de l'intelligence. Elle fonctionne par apprentissage (supervisé ou non) à partir de données. Cet apprentissage consiste à dégager, des données, des catégories et des relations, sans recherche d'extrapolation conceptuelle ou formelle. Cette catégorie d'IA relève du *machine learning* dont le *deep learning*, issu du connexionnisme, au succès actuel fulgurant et qui désigne essentiellement ce que nous entendons maintenant par le terme d'intelligence artificielle. Ses unités fondamentales sont des neurones formels, d'inspiration analogique initiale neuro-mimétique humaine. Elle prend la forme d'un réseau de neurones logiques, c'est-à-dire d'une architecture de combinaison d'opérateurs mathématiques. Elle génère, quant à elle, des espaces vectoriels émergeants dont l'interprétabilité n'est pas immédiatement accessible à l'être humain.

2 Hybridation, interprétation des concepts synthétiques et intrication cognitive

2.1 Hybridation

L'intelligence artificielle hybride (Jiang et al., 2020 ; Gibaut et al., 2023 ; Renkhoff et al., 2024 ; Punzi et al., 2024) se révèle actuellement être une intense et prometteuse voie médiane de recherche en matière l'IA robuste (Alshmrany, 2024 ; Punzi, 2024) et vise à pallier les limites respectives des deux approches de l'intelligence artificielle : concernant l'IA symbolique, son caractère aprioriste, éthéré, son manque d'ancrage matériel ; concernant l'IA empirique, son manque de fiabilité (hallucinations, biais cognitifs mais aussi culturels sources d'iniquité, faible explicabilité) (Du et al., 2023 ; Kandpal et al., 2023 ; McKenna et al., 2023 ; Echterhoff, 2024 ; Kheya, 2024 ; Luo et al., 2024), son coût excessif (humain, financier, énergétique) d'entraînement, ses difficultés avec des tâches spécifiques et son besoin conséquent de réentrainement (partiel) constant.

L'intelligence artificielle neuro-symbolique cherche dès lors à exploiter leurs valeurs ajoutées propres : respect des cadres formels, rigueur, garde-fous, interprétabilité, aisance de construction pour l'IA symbolique, enracinement terrain pour l'IA symbolique. Elle tend donc à combiner logique *a priori* et données pratiques, respects de règles formelles et intégration des informations *hic et nunc*. Elle constitue un système adaptatif, au sens de Piaget (1975), dans la mesure où elle équilibre assimilation (respect et continué de cadres formels) et accommodation (ajustement à la contingence, à la nouveauté, à l'imprévu).

2.2 Interprétation des concepts synthétiques et interfaçage neuro-symbolique

Notre capacité à, en partie, interpréter (Zhao, 2024) les « dimensions d'espace vectoriel » (« catégories de pensée » ou « types de caractéristiques » ou « concepts » synthétiques) (Clark et al., 2019 ; Jawahar et al., 2019 ; Bills et al., 2023 ;

Bricken 2023) auxquelles réagissent respectivement les neurones ou assemblages de neurones des différentes couches d'un réseau de neurones pourrait constituer une clé fondamentale de notre capacité, à un niveau micro-analytique, à authentiquement combiner les aspects symboliques et empiriques de l'intelligence artificielle. En effet, et pour le dire à travers le prisme de la logique formelle, une articulation à granularité fine (Pichat, 2024a, 2024b) du symbolique et de l'empirique nécessite une coordination des arguments et prédicats de la composante symbolique à ceux de la composante empirique, qu'il s'agit dès lors de pouvoir interpréter.

Cette démarche de construction interprétative, que l'on pourrait assimiler à celles de la psychologie culturelle et du relativisme linguistique (Santos, 2024), doit se prémunir de (trop de) biais d'anthropocentrisme (Pichat et al., 2024a, 2024b). C'est-à-dire s'efforcer méthodologiquement de se positionner, autant que possible, dans un processus d'investigation « *culture free* » ; en ne présupposant pas, par exemple, que les concepts synthétiques impliqués seraient nécessairement sémantiquement analogues à nos concepts humains et donc en s'intéressant au prometteur domaine de ce que l'on nomme métaphoriquement les « *alien concepts* » (Bills et al., 2023) (i.e. des concepts associés à des catégories et prédicats qui n'existent pas dans notre culture humaine et pour lesquels nous n'avons pas de terme permettant de les dénoter directement). Ou encore en ne posant pas comme *a priori* que ces concepts synthétiques seraient homogènes comme le sont les nôtres, mais en admettant qu'ils puissent être composites, c'est-à-dire constitués de catégories et prédicats hétérogènes, de différentes natures (Fan et al., 2023 ; Bricken et al., 2023). Ou enfin en ne décrétant pas d'emblée que ces concepts synthétiques seraient unifiés à l'instar des nôtres mais qu'ils puissent être des « champs conceptuels » (Vergnaud, 2009, 2016) distribués simultanément sur plusieurs catégories et prédicats (homogènes ou hétérogènes).

Une telle démarche d'interprétation des concepts synthétiques permettrait de prolonger, en capitalisant sur elles, les conceptions épistémologiques actuelles de l'hybridation de l'intelligence artificielle (Hemmer, 2021). Une de ces passionnantes épistémologies consiste à juxtaposer, à un niveau macro et extrinsèque, le symbolique et l'empirique. Cela en faisant « post-traiter », par un système symbolique, les sorties d'un modèle empirique (ou réciproquement) (Sun, 2024) ; ou en mobilisant des procédés de prompt engineering (Trad, 2024). Une autre de ces épistémologies se traduit par un embodiment du symbolique dans l'empirique ; cela à travers le guidage de l'entraînement d'un modèle empirique par un système symbolique ; en injectant dans la fonction de perte des termes de régularisation pilotés par un dispositif symbolique, à l'instar des « neurones informés par la physique » ; ou en faisant générer par un système symbolique des données synthétiques d'entraînement imprégnées des règles symboliques visées. Mais nous pouvons aller plus loin.

2.3 Intrication cognitive

Par-delà les épistémologies d'immiscibilité cognitive, accéder, dans une certaine mesure, à la « sémantique » des concepts synthétiques neuronaux (portés par les neurones formels ou certains de leurs assemblages) pourrait nous permettre de prétendre à une réelle intrication cognitive cette fois-ci « nanométrique » des registres respectivement empiriques et symboliques de l'intelligence artificielle hybride. Car un mapping (à terme autonome et adaptatif) de la distribution « sémantique » des concepts synthétiques d'un réseau de neurones rendrait possible une connectivité (à terme dynamique, autonome et adaptative) « fine » entre dimensions sémantiques, respectivement empiriques et symboliques, appariables ; et dès lors une coactivité (Clot, 2015) cognitive en elles, à l'instar des réseaux de neurones à mémoire externe dont la mémoire est constituée de connaissances symboliques reformatées. Coactivité qui serait médiatisée par un processus d'interfaçage communicationnel bidirectionnel de transduction d'information entre unités ou modules empiriques et symboliques. Cet interfaçage ayant par exemple (i) comme substrat des embeddings (les connaissances symboliques peuvent par exemple être traduites en embeddings qui peuvent être insérés à l'entrée de couches de neurones), et (ii) comme *modus operandi* l'alignement mathématique des espaces vectoriels des embeddings « symboliques » et des matrices neuronales lorsqu'ils sont initialement hétérogènes et incompatibles.

L'intrication hybride des sphères empiriques et symboliques d'un système d'intelligence artificielle rendrait possible une interaction constante, à fin niveau de granularité, entre connaissances empiriques et symboliques ; cela en permettant à ces deux types d'entités cognitives de s'enrichir mutuellement, dans une dynamique à comprendre telle que celle théorisée par Vygostki (1934) relativement aux notions de concepts spontanés et scientifiques. Les concepts spontanés indique l'auteur sont acquis dans l'expérience pratique, ils sont de nature purement empirique et contextualisée et ne sont pas formalisés ou théorisés, ils demeurent implicites. Les concepts scientifiques sont quant à eux systématisés, explicites, généralisables et abstraits ; ils sont accompagnés de définitions et de règles formelles. Ces deux types de concepts se développent en sens inverse et de façon dialectique au sein d'une interaction constante nous indique Vygostki : les concepts quotidiens développent les concepts scientifiques en les contextualisant et en les ancrant ainsi dans du sens empirique ; à l'inverse, les concepts scientifiques permettent aux concepts spontanés de se généraliser et de s'articuler à d'autres au sein de systèmes plus formels. Ce qui, comme énoncé précédemment, est non seulement de nature (i) à pallier les limites respectives des intelligences empiriques et symboliques, mais également (ii) à transformer la faiblesse de l'une en force de l'autre.

3 Réflexions épistémologiques sur le statut de la cognition synthétique

Cheminer en direction d'une authentique intrication cognitive neuro-symbolique à fine granularité, telle que mentionnée ci-avant, nécessite de nous clarifier plus avant sur le statut épistémologique à accorder à la cognition synthétique à laquelle il s'agit de phaser des éléments, symboliques, relevant de la cognition humaine. Cela afin de mieux comprendre les particularités et les éléments de singularité de cette cognition artificielle, dont une approche trop anthropomorphisée serait de nature à nous empêcher de bien paramétrer l'interfaçage cognitif évoqué. Autrement dit, pour ne pas nous précipiter dans les *a priori* autocentrés de nos catégories de pensée toutes humaines, il s'agit de prendre le temps de nous interroger, dans une démarche épistémologique, sur la « nature » des contenus et processus cognitifs propres aux réseaux de neurones artificiels ; cela, afin de mieux penser les termes d'un possible couplage (au sens de Varela, 1988) entres cognitions neuro-artificielles et symbolico-humaines.

Pour ce faire, nous mobilisons ici les réflexions du constructivisme, dans ses approches systémique et énactive, pour nous interroger sur le statut épistémologique qu'il est pertinent d'attribuer à la cognition synthétique en général, et aux catégories synthétiques en particulier (Pichat 2023, 2024b ; Pichat et al., 2024a, 2024b) ; en nous interrogeant entre autres sur les rapports heuristiques entre psychologie humaine et psychologie des neurones artificiels, et les possibilités comme les limites de la projection de la première sur la seconde. Cela notamment afin de mettre en lumière l'incommensurabilité paradigmatique partielle et le caractère construit de la cognition artificielle, et ainsi de nous prémunir d'un risque trop fort d'anthropomorphisme comme de réalisme cognitif à son endroit.

3.1 La genèse de la singularité catégorielle des neurones formels

La singularité catégorielle (Bills et al., 2023 ; Fan et al., 2023 ; Bricken et al., 2023 ; Zhao et al., 2024) des neurones formels est le fruit de l'ajustement progressif des poids neuronaux (parmi d'autres paramètres) durant la phase d'entraînement du réseau de neurones. En effet, cet apprentissage profond déclenche au sein du réseau de neurones (initialement étalonné avec des poids aléatoires) un processus analogue à ce que Watzlawick (1977) décrivait comme la « confusion » dans le champ de la pensée humaine : une recherche immédiate de significations, de liens, se traduisant par l'accroissement de l'attention et une promptitude à établir des relations, même là où elles pourraient sembler totalement absurdes à un observateur externe ; recherche pouvant s'étendre, nous dit le psychologue constructiviste, jusqu'à inclure des détails tellement petits qu'ils pourraient nous apparaître délirants ; et recherche, complète-t-il, relevant d'une précipitation sur des conclusions étayées par le premier fait tangible qu'on aura cru détecter à travers le brouillard des circonstances. Face à la confusion nous

dit encore Watzlawick, l'individu va chercher et introduire un sens en mobilisant la plus grande ingéniosité pour trouver un ordre à ce qui n'en a aucun, vue de l'extérieur. Et, nous commente enfin l'auteur, lorsque les événements contredisent l'explication (nous pourrions écrire ici la catégorie synthétique fabriquée par un (groupe de) neurone(s) à un stade donné de *deep learning*), une explication encore plus élaborée (i.e. une catégorie synthétique plus complexe) est alors fabriquée.

Ainsi pouvons-nous penser la genèse de la construction cognitive neuronale synthétique vis-à-vis de l'étrangeté dans laquelle elle nous apparaît à nous, observateurs extérieurs anthropocentrés, lorsque nous tentons de la comprendre à des fins de coordination neuro-symbolique : la cognition synthétique est fondamentalement non isomorphe à la nôtre, et naïvement investiguer son explicabilité (Pichat, 2023, 2024a, 2024b) est pour partie une illusion.

3.2 Les catégories synthétiques ne copient pas un monde pré-donné

Les catégories instanciées par les neurones ou par leurs regroupements ne représentent certainement pas une réalité intrinsèque, ontologique, *per se*. Varela (1988) nous met en garde contre le préjugé que le monde tel qu'il est perçu est indépendant du système qui le perçoit : nous présumons, dit-il, d'un monde prédéfini, c'est-à-dire de propriétés qui sont définies préalablement à toute activité cognitive en son endroit ; critiquant épistémologiquement la notion même de représentation, le penseur du Collègue de France argue que la connaissance n'est pas le miroir de la nature ; mais que, au contraire, un système cognitif (ici synthétique) dégage des régularités et des interprétations qui sont fonctions des caractéristiques de ses activités propres (Varela, 1984). Cela, en phase avec la théorie de l'encodage indifférencié qu'il porte avec Maturana dans le champ neurobiologique et qui affirme que la réponse d'une cellule nerveuse n'encode pas la nature physique des *stimuli* : le quoi n'est pas encodé mais seulement le combien (des fonctions d'agrégation et d'activation dirions-nous dans le domaine synthétique) : il n'y a pas de transmission d'information dit avec eux Von Foerster (1984).

Contrairement à l'épistémologie cognitiviste et même connexionniste, pour transposer les propos de Varela (1988), les concepts et processus cognitifs synthétiques ne doivent pas être évalués comme décrivant, ou devant décrire, un réel extérieur prédéterminé, pré-donné ; ceci, à l'instar, nous dit le chercheur, du très ancien idéal d'objectivité, conçu comme l'élimination progressive de l'erreur en vue de l'adéquation toujours plus grande avec une chimérique réalité. Penser une possible intrication neuro-symbolique, c'est dès lors prendre la mesure de fait que la cognition synthétique n'est pas une description, éventuellement subtile, des objets du monde ; mais une reconstruction interprétative singulière au sein de catégories d'analyse dimensionnelle qui sont totalement relatives aux modalités internes spécifiques et aux finalités propres de la cognition artificielle.

3.3 Catégories synthétiques et construction des objets du monde

Dans la continuité immédiate de ce que nous venons de mentionner, les catégories de la cognition synthétique portées par les neurones formels ou leurs assemblages peuvent être pensées comme des équivalents artificiels de ce que Watzlawick (1977) appelle la réalité de niveau 2. C'est-à-dire des construits d'un ordonnancement, d'une ponctuation (nous pourrions écrire ici d'une segmentation catégorielle) de séquences informationnelles. Réalité de niveau 2 qui traduit, nous indique le systémicien dans le domaine de la cognition humaine, la valeur ou la signification construite que l'on assigne à un événement (un token, ou plutôt son expression vectorielle, dans le cadre d'un modèle de langage).

Ces constructions catégorielles créent des réalités (ici conceptuelles synthétiques) différentes écrit Watzlawick vis-à-vis de la pensée humaine, en précisant que ce ne sont pas les événements eux-mêmes que l'on voit différemment en fonction d'elles, mais leur signification présumée (i.e. la catégorie à laquelle on les assigne). Réalité de niveau 2 qui permet de constituer des objets signifiants à partir d'un flux informe de données (Varela, 1988). Notons que même un embedding de départ, avant même tout traitement neuronal, est déjà *de facto* une réalité de niveau 2, étant donné qu'il est déjà une expression vectorielle au sein d'une base dimensionnelle singulière ; à l'instar de ce qu'affirme Von Foerster (1984) en précisant que la réalité de niveau 1 ne peut qu'être très élémentaire et que très vite on passe à de la réalité de niveau 2. Coordonner les cognitions humaines et artificielles, ce n'est donc pas aligner les dernières sur les premières, mais inventer des modalités originales d'interfaçage cognitif de deux mondes catégoriellement pour partie fondamentalement différents.

3.4 Le neurone comme micro-monde catégoriel

Chaque neurone synthétique est un système cognitif propre fruit de son propre « *techno-umwelt* » (Efimov et al., 2023), une unité cognitive locale et distincte de traitement de l'information, avec ses entrées, son activité calculatoire et sa sortie spécifiques. En ce qui concerne ses entrées, ainsi que le rappelaient Von Foerster (1984) et Varela (1988), chaque neurone (ici biologique) ne réagit qu'a son seul environnement immédiat, il ne fonctionne qu'avec son environnement local. Cette clôture informationnelle, au sens de Varela, est l'analogue neuronal de ce que l'auteur signifie lorsqu'il indique, concernant la cognition humaine, qu'il n'existe pour nous qu'un seul monde, celui dont nous faisons l'expérience par nos processus physiologiques (Varela, 1984) ; de même pour le neurone formel, il n'existe qu'un seul univers, celui de son espace vectoriel d'entrée. En ce qui concerne son activité calculatoire, un neurone formel est fondamentalement, i.e. mathématiquement, un vecteur de poids, c'est-à-dire une fonction d'agrégation (couplé, en sa sortie, à une fonction d'activation).

Cette fonction d'agrégation (cf graphe ci-après), de la forme $\sum(w_{i,j}x_{i,j})+a$, est l'opérateur *princeps* de la construction catégorielle singulière réalisée *hic et nunc* par ce neurone ; car cette fonction d'agrégation préside à la particular-

ité de la coordination catégorielle pondérée qui va être réalisée des segments catégoriels d'entrée qui sont les siens, c'est-à-dire des dimensions catégorielles de son espace vectoriel d'entrée (Pichat, 2024b). Et suite à ce traitement catégoriel, une sortie catégorielle exclusive est créée. Cette nouvelle segmentation catégorielle résultante est une construction originale locale qui est propre à ce neurone (i.e. à sa fonction d'agrégation puis d'activation). Comme le mentionne Watzlawick (1977) dans le champ systémique de la pensée humaine : le langage (nous pourrions écrire ici une coordonnée d'embedding sortant) ne se contente pas de transmettre des informations mais exprime en même temps une vision du monde ; de façon transposée, le neurone formel ne transmet pas (en sortie) passivement une information préexistante neutre mais un angle de vue catégoriel *sui generis*, qu'il a authentiquement élaboré.

Chercher à intriquer les activités cognitives humaines et synthétiques neuronales implique de comprendre qu'un réseau de neurones relève fondamentalement d'une cognition catégorielle, très différente des projections anthropocentrées de nos propres modes de raisonnement humains (de phénoménologies apparentes) beaucoup plus complexes et diversifiés ; et que c'est certainement au niveau de ces catégories synthétiques singulières qu'il est pertinent de positionner une intrication neuro-symbolique.

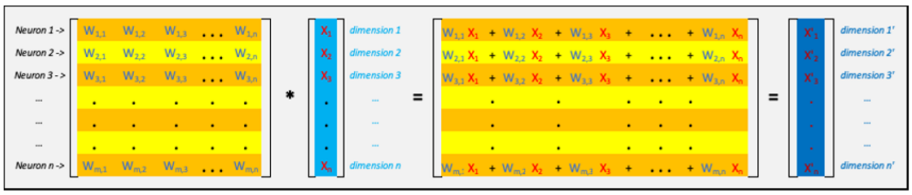

Figure 1: Fonction d'aggrégation

3.5 L'incorporation mathématique et architecturale des catégories synthétiques

Ainsi que nous le donne notamment à voir le graphe ci-avant relatif à la fonction d'agrégation neuronale, les catégories synthétiques ne sont pas des entités transcendantes *ex nihilo*, mais le fruit, dans une approche de cognition incarnée, de choix humains mathématiques, architecturaux et d'entraînement qui ont présidé à la fabrication du système neuronal au sein duquel elles émergent. En transposant Watzlawick (1977) à leur endroit, nous pourrions en effet écrire que l'on s'émerveille du mirage catégoriel de la cognition artificielle que nous avons pourtant nous-même généré en première instance ; ou, comme le dit Varela (1984), nous (observateurs humains) percevons le monde des catégories synthétiques sans nous rendre compte de tout ce que nous faisons pour le trouver ainsi : les segmentations catégorielles observées sont le résultat des activités des fonctions d'agrégation, des fonctions d'activation et autres paramètres neuronaux choisis, de nature à générer des « *techno-umwelt* » (Efimov et al., 2023) spécifiques.

Comme le précise Von Foerster (1984) concernant le système neuronal humain, la nature et l'organisation des opérations neuronales, ici synthétiques, vont générer le calcul et la construction de certaines abstractions, de différents encodages catégoriels. Les catégories synthétiques sont dès lors la résultante d'un processus d'incorporation mathématique et d'embodiement architectural. Dans ce registre, Von Foerster, à nouveau, indique que ces propriétés catégorielles n'existent que dans les opérations et les organisations de ces opérations qui ont présidé à leur construction ; et certainement, *a fortiori*, pour le cas des couches profondes. Ou, reformulé par Varela (1988), les opérations de la pensée synthétique sont contraintes sémantiquement car toutes les distinctions catégorielles que nous pouvons identifier ont été générées par le programmeur de ce système artificiel : la base d'embeddings choisie, le type de données d'entraînement, la nature des feedbacks administrés durant l'entraînement, etc.

Créer des *modi operandi* d'intrication neuro-symbolique nous invite dès lors à inventer des systèmes de traduction, de transduction des informations symboliques en des modes informationnels synthétiques qui soient adaptés au caractère mathématiquement incorporé de la cognition neuronale synthétique ; par exemple en reconstruisant ces connaissances symboliques à « communiquer » ou à « injecter » dans des dimensions catégorielles d'espaces vectoriels « lisibles » par le formatage vectoriel d'une couche neuronale donnée.

3.6 Le processus de l'émergence catégorielle synthétique

L'interprétation des catégories neuronales synthétiques implique potentiellement une approche structurale. En effet, ainsi que l'explicite Varela (1988) concernant les neurones biologiques, ces derniers doivent être étudiés en tant que membres d'ensembles neuronaux de différents types d'ampleur, qui, écrit-il, apparaissent et disparaissent régulièrement au gré du fil de leurs interactions coopératives.

Dans ce registre, à titre d'exemple, nous nous interrogeons actuellement sur la pertinence d'étudier la genèse des catégories neuronales artificielles sur la base des catégories de leurs neurones précurseurs respectifs, à travers des processus de phasage et de déphasage catégoriel entre ces neurones précurseurs (ou plutôt entre certaines de leurs sous-dimensions catégorielles extraites), selon les tokens impliqués ; processus portés par la fonction d'agrégation (cf graphe précédent) qui définit la catégorie d'un neurone donné d'une couche n comme la coordination pondérée des catégories propres à chacun de ses neurones précurseurs contributifs de la couche n-1. Autrement dit, nous nous demandons comment comprendre, pour un neurone donné, la genèse de sa segmentation catégorielle de sortie (sa dimension vectorielle de sortie) comme résultante de l'agrégation de ses segments catégoriels d'entrée (ou plutôt des sous-dimensions de son espace vectoriel d'entrée) ; cela, à travers un processus de projection vectorielle interne se traduisant par d'originales « reconstructions catégorielles » (des réductions catégorielles par intersection catégorielle, des extensions catégorielles par union catégorielle).

Ainsi que nous le livre Varela (1988), la configuration du système neuronal, dans notre présent exemple le vecteur de poids de la fonction d'agrégation,

permet l'émergence d'une coopération neuronale lorsque les états de chaque neurone (toujours dans notre exemple, précurseur) impliqué atteignent un certain stade (ici l'activation des neurones précurseurs par le token en jeu). Telle est la manière dont pourrait peut-être se comprendre en partie une catégorie neuronale synthétique, dans la logique d'autonomie et d'auto-organisation de Varela (1984) : une unité (ici la catégorie synthétique portée par un neurone) se dégage d'elle-même par phasage, au sein de la fonction d'agrégation, d'entités neuronales (d'entrée) qui se coordonnent dynamiquement pour produire une nouvelle dimension (en sortie) de segmentation catégorielle plus abstraite.

Mieux comprendre ces dynamiques d'interactions neuronales sur la genèse des contenus (dont les catégories) et processus cognitifs des systèmes neuronaux artificiels, ainsi que le font par exemple Bricken et al.(2023) dans un autre registre, est de nature à constituer une précieuse avancée en termes de possibilités d'interfaçage neuro-symbolique ; cela, par exemple, en nous permettant de nous clarifier plus avant quant aux *topoï* (lieux) neuronaux (et à leur niveau d'ampleur, neurones ou assemblages de neurones) précis au sein desquels il peut être pertinent d'injecter des paramètres symboliques, dans la mesure où ces lieux seraient spécifiquement impliqués dans des catégories synthétiques liées ou liables aux catégories de pensée des connaissances symboliques en jeu.

3.7 Paradoxe des catégories synthétiques et processus de couplage

Les catégories synthétiques vectorisées par les neurones artificiels, ainsi que le montre l'observation empirique des tokens pour lesquels ces neurones s'activent (Clark et al., 2019 ; Jawahar et al., 2019 ; Bricken et al., 2023), comportent résolument des dimensions liées à des catégories de pensée humaines ; et en même temps, ces catégories semblent sémantiquement d'une hétérogénéité de sous-dimensions catégorielles dont l'être humain ne peut comprendre la logique unificatrice sous-jacente (Bills et al., 2023). C'est le paradoxe apparent des catégories synthétiques, mi-humaines et mi-étrangères à notre mode de segmentation du monde.

La notion de couplage structurel telle que développée par Varela (1988) constitue potentiellement un repère pour comprendre cette dualité qui pourrait nous sembler phénoménologiquement contradictoire. En effet, d'un côté, les données d'entraînement (phrases) d'un modèle de langage tout comme les feedbacks qui lui sont administrés (en fonction de la conformité de ses réponses à celles, humainement sémantiquement adaptées, qui sont attendues) relèvent pleinement d'un registre humain de signification, i.e. sont fondés sur des catégories humaines de pensée. Et, d'un autre côté, l'auto-ajustement progressif de ses poids neuronaux, par le modèle durant son apprentissage, lui fait fabriquer des segmentations catégorielles qui lui sont propres (alien concepts) et qui, sans recherche de respect de la sémantique humaine, lui permettent « coûte que coûte » de faire montre *in fine* d'une réponse qui est celle qui est attendue par le régulateur humain.

Dans la lignée immédiatement de ce que nous venons d'écrire, Varela définit

précisément la cognition comme l'historique (ici durant la phase de *deep learning*) du couplage structurel progressif entre (i) le monde interne (ou plutôt les différents éléments interconnectés du monde interne) d'un système cognitif (capable de subir des changements structuraux au cours de son histoire, ici des changements de ses paramètres) et (ii) des caractéristiques d'un monde extérieur auquel il doit se coordonner (ici, la sémantique humaine) ; couplage qui enacte (fait émerger), poursuit Varela, un monde de signification à la croisée de contraintes externes et internes. Ce monde de signification, dans le cas de la cognition synthétique, est composé des catégories singulières de cette cognition synthétique, catégories qui sont des régularités autant extraites que construites, permettant, nous dit enfin Varela, une articulation significative à un monde (extérieur) indéfini, une adjonction à un monde de signification (en continuelle évolution) ou la création d'un nouveau monde de signification, par modelage mutuel d'un monde commun au moyen d'une action conjuguée.

La notion de couplage nous apparaît comme une notion clé pour penser épistémologiquement la nature de l'intrication cognitive neuro-symbolique qui devrait dès lors consister en des catégories mixtes, synthétiques et humaines, où s'articulent, se rencontrent des univers sémantiques différents mais localement convergeant. Cela, par exemple, et pour reprendre Vygostki (1934), en construisant des embeddings (adaptés localement à l'espace vectoriel interne correspondant à leur lieu d'implantation neuronal) de connaissances ou de critères symboliques à respecter, qui vont ancrer ces éléments symboliques « théoriques » (ces concepts scientifiques dirait Vygoski) dans des dimensions catégorielles empiriques (des concepts pratiques dirait Vygotski) accessibles à la cognition synthétique ; et, inversement, embeddings qui vont permettre de généraliser, de formaliser et d'aligner plus avant avec les attendus normatifs humains les catégories-en-acte (ainsi que le mentionnerait Vergnaud, 2009) incorporées dans les (assemblages de) neurones formels des zones neuronales spécifiquement impliquées dans les catégories relevant des éléments symboliques à injecter.

4 Conclusion : les IA générales de prochaine génération

Une valeur ajoutée de l'intrication cognitive neuro-symbolique pensée dans une logique épistémologique constructiviste, comme nous venons de tenter de commencer à le faire, est de faciliter *de facto* une fabrication d'un nouveau type, peut-être moins naïf, d'explicabilité (Budding, 2024) et donc de diagnostic, de débogage et d'optimisation de systèmes d'intelligence artificielle en leurs composantes connexionnistes. Mais plus encore, un autre intérêt de l'intrication cognitive est qu'elle facilite potentiellement une série d'améliorations que les sciences cognitives et les neurosciences (Minsky, 1988 ; Sun, 2023 ; Xie, 2023) ainsi que la réflexion philosophique en matière d'IA (Efimov et al., 2021, 2023) nous invitent à apporter à nos systèmes d'intelligence artificielle actuels pour les faire évoluer vers des intelligences artificielles générales qualitativement beaucoup

plus abouties, par-delà l'illusion quantitative du « *more data* » : (i) modularité, (ii) connectivité (temporaire ou stable) entre unités et modules, (iii) flexibilité des contenus, processus et des structures qui les coordonnent comme des connexions qui les relient, (iv) coordination de processus top-down et bottom-up, (v) circularité, (vi) multimodalité, (vii) métacognition, (viii) auto-modification adaptative et régulation évolutive autonome de ces contenus, processus, structure et connexions. Ces caractéristiques neuro-mimétiques pouvant et devant être d'emblée pensées au sein d'une logique duelle, dialectique d'intrication neuro-symbolique systémique et énactive.

Remerciements

L'auteur tient à remercier Jourdan Wilson et William Pogrund pour leur travail méticuleux de révision et de mise en page de cet article.

Bibliographie

[1] Alshmrany, K. M., Aldughaim, M., Wei, C., Sweet, T., Allmendinger, R., & Cordeiro, L. C. (2024). *FuSeBMC AI: Acceleration of Hybrid Approach through Machine Learning*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.06031

[2] Bills, S., Cammarata, N., Mossing, D., Saunders, W., Wu, J., Tillman, H., Gao, L., Goh, G., Sutskever, I., & Leike, J. (2023). *Language models can explain neurons in language models*. OpenAI. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

[3] Bricken, T., Schaeffer, R., Olshausen, B., & Kreiman, G. (2023). *Emergence of Sparse Representations from Noise*. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research, 202, 3148-3191. https://proceedings.mlr.press/v202/bricken23a.html

[4] Budding, C., et al. (2024). *Does Explainable AI Need Cognitive Models?* Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46).

[5] Clot, Y. (2015). *La fonction psychologique du travail*. Paris: PUF.

[6] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look At? An Analysis of BERT's Attention*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1906.04341

[7] Du, Y., Konyushkova, K., Denil, M., Raju, A., Landon, J., Hill, F., Nando, D. F., & Cabi, S. (2023). *Vision-Language Models as Success Detectors*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.07280

[8] Echterhoff, J., Yan, A., Han, K., Abdelraouf, A., Gupta, R., & McAuley, J. (2024). *Driving through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving*. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv57701.2024.00718

[9] Efimov, A., Dubrovsky, D., & Matveev, F. (2023). *What's Stopping Us Achieving Artificial General Intelligence?* Philosophy Now, April/May.

[10] Efimov, A., Dubrovsky, D., & Turchin, N. (2021). *Walking Through the Turing Wall*. Proceedings of the 20th IFAC Conference on Technology, Culture, and International Stability TECIS 2021: Moscow, Russian Federation, 14–17 September 2021.

[11] Fan, Y., Dalvi, F., Durrani, N., & Sajjad, H. (2023). *Evaluating Neuron Interpretation Methods of NLP Models*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.12608

[12] Gibaut, W., Pereira, L., Grassiotto, F., Osorio, A., Gadioli, E., Munoz, A., Gomes, S., & Santos, C. D. (2023). *Neurosymbolic AI and its Taxonomy: A Survey*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.08876

[13] Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). *Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review*. Pacific Asia Conference on Information Systems, 78. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1077&context=pacis2021

[14] Jawahar, G., Sagot, B., & Seddah, D. (2019). *What Does BERT Learn about the Structure of Language?* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1356

[15] Jiang, K., Wang, W., Wang, A., & Wu, H. (2020). *Network Intrusion Detection Combined Hybrid Sampling with Deep Hierarchical Network*. IEEE Access, 8, 32464-32476. https://doi.org/10.1109/access.2020.2973730

[16] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). *Large Language Models Struggle to Learn Long-Tail Knowledge*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.08411

[17] Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). *The Pursuit of Fairness in Artificial Intelligence Models: A Survey*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.17333

[18] Luo, J., Zhuo, W., Liu, S., & Xu, B. (2024). *The Optimization of Carbon Emission Prediction in Low Carbon Energy Economy under Big Data*. IEEE Access, 12, 14690-14702. https://doi.org/10.1109/access.2024.3351468

[19] McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023). *Sources of Hallucination by Large Language Models on Inference Tasks*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.14552

[20] Minsky, M. (1988). *The Society of Mind*. New York: Simon & Schuster.

[21] Piaget, J. (1975). *Comments on Mathematical Education*. In Cambridge University Press eBooks (pp. 79-87). https://doi.org/10.1017/cbo9781139013536.004

[22] Pichat, M. (2023). *Collaboration des intelligences humaine et artificielle: alignement et psychologie de l'IA. Actes du colloque « Intelligence artificielle collaborative & impacts managériaux au sein des organisations » du 30/06/2023 coorganisé par l'Université Paris Dauphine-PSL et le Cabinet Chrysippe R&D.* `https://www.youtube.com/watch?v=kG9Uv8-7OyQ&list=PLD25p-Bh6_swAk-TrFgk4lIQ6MQ2r5NTv&index=3`

[23] Pichat, M. (2024a). *Psychology of Artificial Intelligence: Epistemological Markers of the Cognitive Analysis of Neural Networks*. arXiv (Cornell University). https://arxiv.org/abs/2407.09563

[24] Pichat, M., Campoli, E., Pogrund, W., Wilson, J., Veilllet-Guillem, M., Melkozerov, A., Pichat, P., Gasparian, A., Poumay, J., & Demarchi, S. (2024b). *Neuropsychology of AI: Relationship Between Activation Proximity and Categorical Proximity Within Neural Categories of Synthetic Cognition*. arXiv (Cornell University). https://arxiv.org/abs/2410.11868

[25] Pichat, M., Campoli, E., Pogrund, W., Wilson, J., Veilllet-Guillem, M., Melkozerov, A., Pichat, P., Gasparian, A., Poumay, J., & Demarchi, S. (2024c). *Neuropsychology and Explainability of AI: A Distributional Approach to the Relationship Between Activation & Similarity of Neural Categories in Synthetic Cognition*. arXiv (Cornell University).https://arxiv.org/abs/2411.07243

[26] Punzi, C., Pellungrini, R., Setzu, M., Giannotti, F., & Pedreschi, D. (2024). *AI, Meet Human: Learning Paradigms for Hybrid Decision Making Systems*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.06287

[27] Renkhoff, J., Feng, K., Meier-Doernberg, M., Velasquez, A., & Song, H. H. (2024). *A Survey on Verification and Validation, Testing and Evaluations of Neurosymbolic Artificial Intelligence*. IEEE Transactions on Artificial Intelligence, 5(8), 3765-3779. https://doi.org/10.1109/tai.2024.3351798

[28] Santos, J. A., Massolo, A., & Durante, S. (2024). *Logical Pluralism and Linguistic Relativism*. Filosofia Unisinos, 25(2), 1-10. https://doi.org/10.4013/fsu.2024.252.04

[29] Sun, D., Wang, H., & Xiong, J. (2023). *Would You Like to Listen to My Music, My Friend? An Experiment on AI Musicians*. International Journal of Human-Computer Interaction, 40(12), 3133-3143. https://doi.org/10.1080/10447318.2023.2181872

[30] Sun, R. (2024). *Can a Cognitive Architecture Fundamentally Enhance LLMs? Or Vice Versa?* https://arxiv.org/abs/2401.10444

[31] Trad, F., & Chehab, A. (2024). *Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models*. Machine Learning and Knowledge Extraction, 6(1), 367-384. https://doi.org/10.3390/make6010018

[32] Vergnaud, G. (2009). *The Theory of Conceptual Fields*. Human Development, 52(2), 83-94. https://doi.org/10.1159/000202727

[33] Vergnaud, G. (2016). *The Nature of Mathematical Concepts*. In *Learning and Teaching Mathematics* (pp. 5-28). Psychology Press.

[34] Varela, F. (1988). *Cognitive Science: A Cartography of Current Ideas*. New York/Leuven: Pergamon Press/Leuven University Press.

[35] Varela, F. (1984). *The Creative Circle*. In P. Watzlawick (Ed.), *The Invented Reality*. London: W. W. Norton & Co.

[36] Von Foerster, H. (1984). *Construction of a Reality*. In P. Watzlawick (Ed.), *The Invented Reality*. London: W. W. Norton & Co.

[37] Vygotsky, L. S. (1934). *Thought and Language*. MIT Press.

[38] Watzlawick, P. (1977). *How Real is Real?* London: Vintage Books.

[39] Xie, H. (2023). *The Promising Future of Cognitive Science and Artificial Intelligence*. Nature Reviews Psychology, 2(4), 202-202.

[40] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). *Explainability for Large Language Models: A Survey*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.01029