COMMENT PENSENT LES IA? LES 3 FACTEURS MATHEMATICO-COGNITIFS DE LA SEGMENTATION CATEGORIELLE OPEREE PAR LES NEURONES SYNTHETIQUES

Michael Pichat*
William Pogrund†
Armanush Gasparian‡
Paloma Pichat§
Samuel Demarchi¶
Michael Veillet-Guillem‖

Résumé

Comment les neurones synthétiques des modèles de langage créent des "catégories de pensée" pour segmenter et analyser l'environnement informationnel qui est le leur ?
Quelles sont les caractéristiques cognitives, au niveau même des neurones formels, de cette pensée artificielle catégorielle ? En nous fondant épistémologiquement sur la nature mathématique des opérations algébriques portées par les fonctions d'agrégation neuronale, nous tentons de mettre à jour des facteurs mathématico-cognitifs qui façonnent génétiquement la reconstruction catégorielle du monde informationnel qui lui fait face par la cognition artificielle. Cela à travers les concepts d'amorçage, d'attention et de phasage catégoriels.

*Neocognition (Chrysippe R&D) Université de Paris & Facultés Libres de Philosophie et de Psychologie de Paris (ER IPC)
†Neocognition and NP - Phelma, Université Grenoble Alpes
‡Neocognition
§Neocognition et Faculté de Médecine de Lyon Est (Université Lyon 1)
¶Neocognition (Chrysippe R&D), Université Paris 8 & Facultés Libres de Philosophie et de Psychologie de Paris
‖Neocognition (Chrysippe R&D) et Epitech Paris

1 Introduction

1.1 Explicabilité synthétique et inférence cognitive

Rendre explicable un réseau de neurones artificiels signifie traduire ses opérations dans un langage accessible et logique pour les humains [25, 55, 56, 57]. Cela implique d'interroger les actions observables du réseau dans un cadre de référence interprétatif qui permette d'assigner un sens pertinent à ses opérations. Dans le cadre propre de notre approche, nous mobilisons pour ce faire le référentiel explicatif de la psychologie cognitive, utilisant les concepts issus de la pensée humaine comme base pour établir des liens heuristiques ou analogiques entre intelligences humaine et artificielle. Démarche devant être réalisée en nous interrogeant en permanence sur le risque d'erreurs consistant à attribuer des traits humains aux algorithmes [51], à confondre comportement et cognition [12], ou à fusionner observateur et système observé, risque mis à jour par la cybernétique, la systémie comme les sciences cognitives de l'énaction [77, 78, 70, 72].

L'utilité pratique de cette démarche d'explicabilité cognitive se déploie en deux axes. Premièrement, elle permet de prévenir les réponses erronées, voire dangereuses, du système neuronal artificiel (Luo et al., 2024), telles que les biais cognitifs [29],ou culturels [42], ainsi que les hallucinations [40, 49], ou l'accent excessif mis sur certaines entrées [26]. Deuxièmement, elle permet d'améliorer l'efficacité des modèles de langage [8] en les alignant plus avant sur les attendus humains [44].

Dans cette étude, nous explorons une approche d'explicabilité focalisée sur une granularité cognitive fine, appelée explicabilité mécaniste. Au lieu d'examiner globalement les outputs d'un réseau de neurones en relation avec ses entrées [85], nous nous concentrons dès lors sur une analyse plus microscopique. Cette démarche se penche ainsi sur les unités cognitives fondamentales des réseaux de neurones formels, les neurones synthétiques, seuls ou en groupes au sein de couches [21, 22, 32, 52]. Plus pécifiquement, notre objectif sera ici de pénétrer et d'inférer le mécanisme cognitif interne des réseaux artificiels dans une dynamique génétique afin de comprendre comment les catégories et concepts vectorisés par les neurones formels sont localement constitués.

2 Statut épistémologique des catégories de pensée synthétiques

Par construction structurelle (i.e. architecturale) et fonctionnelle (i.e. mathématique), la cognition des constituants d'un réseau de neurones synthétiques est pour partie une pensée catégorielle [11, 32, 14, 43, 86, 56, 57]. En effet, de façon schématique, le mode de fonctionnement de chaque neurone formel se caractérise par les trois phases suivantes :

1. **Intégration**: en entrée, des sorties de ses neurones formels précurseurs, chacune de ces sorties pouvant être interprétée comme le niveau

d'appartenance, de l'élément actuellement traité par le réseau (un token donné pour les modèles de langage), à la catégorie associée à un précurseur.

2. **Combinaison additive pondérée**: par une fonction dite d'agrégation,[1]. de ces entrées afin d'élaborer sur cette base une nouvelle catégorie résultante ; catégorie dont le contraste est augmenté par une fonction non linéaire dite d'activation, à des fins de sparsité (Raieli et al., 2024 ; Xiao et al., 2024).

3. **Production d'une sortie**, qui sera à son tour utilisée par les neurones successeurs à venir.

Pour chaque membre (token) d'une catégorie synthétique, une valeur d'activation associée à cet élément pour cette catégorie indique quantitativement, par-delà une logique ensembliste classique d'appartenance dichotomique, à quel point cet élément appartient à cette catégorie artificielle ; en phase avec ce que la logique floue [83, 80]. Pour chaque catégorie, son extension [51], peut alors être définie comme l'ensemble des éléments (tokens) associés à une valeur d'activation strictement positive et éventuellement supérieure à un seuil à arbitrairement fixer, dans le cadre d'une α-coupe au sens de la logique floue.

Dans le positionnement épistémologique qui est le nôtre, les catégories synthétiques, comme humaines [73], ne sont pas des entités cognitives transcendantes mais immanentes. Chaque catégorie synthétique est une authentique construction cognitive dans la mesure où elle est opérée par le réseau de neurones lui-même, durant sa phase d'entrainement. Une catégorie artificielle est ainsi une création de segmentation, une fabrication de dimension dans l'espace, infini et non prédéterminé, de l'ensemble des arguments et prédicats qu'il est possible d'élaborer [51].Ces arguments et prédicats pouvant relever d'éléments analogues à des catégories humaines de pensée existantes (catégories « human like ») ou pas ; dans ce dernier cas on parlera alors métaphoriquement de catégories « alien like », pouvant par exemple correspondre à des construits statistiques [11] ou à des « concepts polysémiques » [14, 52] notion en partie anthropocentrée consistant parfois à curieusement s'étonner que l'univers catégoriel des réseaux de neurones artificiels soit différent de celui des êtres humains.

Concernant la segmentation catégorielle singulière réalisée par un neurone synthétique, la question n'est ainsi pas, dans une logique enactive [70] [71] que nous mobilisons ici, celle de son adéquation ontologique à un prétendu réel préexistant mais celle de sa fonctionnalité (Varela dirait de son couplage) dans le cadre

[1]Notons que Bills et al. (2023), dont nous allons exploiter les donnés dans le cadre de cette présente étude, au sein de leur compte github associé à leur article, indiquent une liste « of the upstream and downstream neurons with the most positive and negative connections » ; liste dont ils donnent la définition opérationnelle suivante : « Definition of connection weights : neuron-neuron: for two neurons $(l1, n1) and (l2, n2) with l1<l2$, the connection strength is defined as $h\{l1\}.mlp.c\_proj.w[:, n1, :]@diag(h\{l2\}.ln\_2.g)@h\{l2\}.mlp.c\_fc.w[:, :, n2]$.. Cette liste définit, dans le cadre des couches denses $i.e. couches pleinement connectes$ de GPT2-XL, les poids à travers lesquels chaque neurone d'une couche d'arrivée n+1 est relié à l'ensemble des neurones de la couche précédente n. C'est sur la base de ces poids que les fonctions linéaires d'agrégations neuronales, auxquelles nous nous référons dans ce présent article, opèrent.

de la tâche finalisée qui est la sienne [7].. Une catégorie est dès lors, dans une perspective constructiviste, avant tout une projection de propriété pragmatiquement utile et fabriquée sur des éléments du monde et non pas une reconnaissance d'une propriété pré-donnée de ces éléments. Les catégories construites par les neurones synthétiques sont dès lors ici conçues comme cognitivement proches de ce que Vergnaud [73, 74] nomme des concepts-en-acte ; c'est-à-dire, des arguments et prédicats tenus pour pertinents (fonctionnels) pour réaliser une tâche mais nullement dénotés (verbalisés), justifiés, théorisés et encore moins conscientisés.

Précisons que ces catégories synthétiques peuvent être inférées à différentes échelles de granularité d'observation d'un réseau de neurones : à celle d'un neurone formel (on parlera alors d'une catégorie neuralement localisée) [11], ou à celle d'une couche voire de connexions inter-couches (on parlera alors d'une catégorie distribuée) [14, 52].

3 Problématique

Comment les neurones synthétiques construisent les dimensions catégorielles à travers lesquelles ils segmentent et analysent le monde (de tokens dans le cas des LLM) qui est le leur ? Quelles sont les caractéristiques développementales de cette pensée artificielle catégorielle et de ces catégories vectorisées par les neurones synthétiques ? Et plus spécifiquement, quels sont les facteurs génétiques qui impactent ou président à ces constructions catégorielles ? Et, encore plus précisément, quels sont les facteurs qui vont piloter le niveau d'appartenance (i.e. le niveau d'activation) d'un token au sein d'une catégorie neuronale synthétique et ainsi déterminer l'extension et donc la définition ou la « sémantique » de cette catégorie ? Autrement dit, comment, quantitativement mais aussi qualitativement, de tels facteurs constituent les variables génétiques de la fonction de segmentation catégorielle (du monde des tokens) opérée par les neurones synthétiques ?

Investiguer ces questions nécessite de prendre acte que les propriétés cognitives et conceptuelles des réseaux de neurones artificiels ne sont pas des émergences apparaissant par magie ou par hasard. Dans une logique de cognition incarnée, incorporée [70, 71, 2, 54] ces propriétés sont le fruit direct des caractéristiques spécifiques du dispositif matériel au sein des paramètres et des contraintes duquel elles émergent.

Une composante centrale, à la fois structurale et fonctionnelle, du dispositif tangible d'un réseau de neurones, est la fonction d'agrégation qui préside à la combinaison linéaire et à la projection vectorielle des dimensions catégorielles en input en une dimension catégorielle résultante en output. Cette fonction d'agrégation, parmi d'autres éléments (dont la fonction d'activation bien entendu), détermine ainsi génétiquement et fonctionnellement le façonnage de la segmentation catégorielle dimensionnelle spécifiquement implémentée par chaque neurone formel.

Une observation de la nature et des opérateurs constitutifs de cette fonc-

tion d'agrégation, de type $\sum(w_{i,j}x_{i,j}) + a$, nous donne assez immédiatement à penser qu'elle génère et formate mathématiquement la segmentation catégorielle réalisée par les neurones synthétiques à travers au moins trois facteurs mathématico-cognitifs. Facteurs que nous allons investiguer dans le cadre de ce travail exploratoire. Un premier facteur est lié à la variable $x_{i,j}$de cette fonction d'agrégation, correspondant aux valeurs d'activation des output catégoriels des neurones précurseurs ; facteur que nous allons cognitivement interpréter en terme d'amorçage catégoriel (ou effet X). Un deuxième facteur est quant à lui lié au paramètre $w_{i,j}$correspondant à la pondération assignée à ces outputs ; nous allons cognitivement l'interpréter en terme d'amorçage catégoriel (ou effet W). Enfin, un dernier facteur est lié à la combinaison linéaire additive Σ des termes $w_{i,j}x_{i,j}$ constitutive de la fonction d'agrégation ; nous le dénoterons cognitivement en terme de phasage catégoriel (ou effet $\sum$).

4 Méthodologie

4.1 Positionnement méthodologique

Pour mieux appréhender le positionnement de notre travail exploratoire, nous présentons ici un aperçu succinct et non exhaustif des diverses approches techniques qui, avec plus ou moins de finesse de granularité cognitive, cherchent à extraire le contenu ou les processus informationnels contenus dans les réseaux de neurones formels, qu'ils soient organisés en couches, en groupes ou en réseau complet. Ces approches ne sont pas mutuellement exclusives et peuvent se croiser en partie.

Comme sommairement mentionné précédemment, les études à spectre macrocognitif se concentrent sur l'analyse des différences entre les entrées et les sorties, visant à comprendre le lien entre les données initiales et les résultats dans un modèle de langage. Parmi cet ensemble, les méthodes basées sur les gradients évaluent le rôle de chaque donnée d'entrée en exploitant les dérivées concernant chaque dimension d'entrée [30]. Les caractéristiques des entrées peuvent être évaluées en fonction d'éléments comme les traits [23], les scores d'importance des tokens [30] ou les poids attribués à l'attention [5]. En parallèle, les approches par exemples cherchent à déterminer comment les sorties varient face à différents inputs, en observant l'effet de légères modifications d'entrée [76] ou d'altérations telles que la suppression, la négation, le mélange ou le masquage des entrées [4, 79, 69].Par ailleurs, certains travaux s'intéressent au mapping conceptuel des entrées pour quantifier leur contribution aux résultats observés [17].

Les méthodes à granularité cognitive plus fine se focalisent sur les états intermédiaires du modèle de langage plutôt que sur sa sortie finale, examinant les sorties ou états internes partiels de neurones ou groupes de neurones. Dans ce contexte, certaines approches analysent et décomposent linéairement le score d'activation d'un neurone d'une certaine couche en rapport avec ses entrées (neurones, têtes d'attention ou tokens) dans la couche précédente [75]. D'autres ten-

dent à simplifier les fonctions d'activation pour en faciliter l'interprétation[76]. De plus, certaines techniques, en se basant sur le vocabulaire du modèle, portent sur l'extraction des connaissances encodées en projetant les connexions et représentations intermédiaires à travers une matrice de correspondance [24, 35]. Enfin, des méthodologies exploitent les statistiques d'activation neuronale en réaction à des ensembles de données [11, 50, 28, 76, 20]. Notre présente étude exploratoire s'insère spécifiquement dans ce dernier ensemble de démarches.

4.2 Options méthodologiques

Pour notre présente recherche exploratoire, nous nous sommes intéressés au modèle GPT proposé par OpenAI, sélectionnant plus particulièrement sa version GPT-2XL. Ce choix porté sur GPT-2XL est dû au fait qu'il est suffisamment complexe pour nous permettre d'examiner des phénomènes cognitifs synthétiques avancés sans atteindre la sophistication de GPT-4, ou encore plus de sa version multimodale GPT-4o. Une autre raison pragmatique a motivé notre préférence pour GPT-2XL : en 2023, OpenAI a partagé, dans l'article de Bills et al. (2023), des détails sur les paramètres ainsi que les valeurs d'activation de ses neurones, informations qui serviront de base à notre analyse.

Dans un souci de simplicité, cette présente étude exploratoire s'est limitée aux deux premières couches de GPT-2XL (couche 0 et 1) comprenant chacune 6400 neurones. Concernant les tokens et leurs valeurs d'activation parmi ces 12800 neurones formels (soit 2 x 6400), nous avons décidé de considérer, pour chaque neurone, les 100 tokens présentant les valeurs d'activation moyennes les plus élevées (nommés « core-tokens »).

4.3 Choix statistiques

Nos analyses statistiques descriptives et inférentielles ont été réalisées au moyen des bibliothèques Python de la suite SciPy, après consultation de Howell[39] et Beaufils[10].

Pour investiguer la normalité de nos données, une condition nécessaire pour l'exécution de tests paramétriques, nous avons adopté une double approche. D'une part, nous avons employé divers tests inférentiels : le test de Shapiro-Wilk (efficace pour de petits échantillons), le test de Lilliefors (adapté aux petits échantillons lorsque les paramètres de la distribution normale sont inconnus et estimés à partir des données), le test de Kolmogorov-Smirnov (préférable pour les grands échantillons), et le test de Jarque-Bera (qui se concentre sur la symétrie et l'aplatissement, valable pour les grands échantillons). D'autre part, nous avons utilisé une approche descriptive avec des indices tels que skewness (symétrie) et kurtosis (aplatissement), et des méthodes graphiques comme le QQ-plot pour comparer la distribution observée à une distribution théorique normale.

Les résultats, non reproduits ici, indiquent une normalité assez relative de nos données, nous orientant alors principalement vers des études de corrélation ordinale de Spearman dans l'analyse des relations entre variables associées à nos

hypothèses, nous permettant ainsi d'éviter les prérequis de normalité et d'éviter les biais introduits par les valeurs atypiques. Au besoin, nous avons utilisé des tests chi^2 univariés d'ajustement pour inférer la significativité des phénomènes observés (notamment quant à la positivité et la significativité des corrélations ordinales obtenues pour chaque neurone de la couche 1).

Dans notre présent cadre statistique, les unités composées utilisées incluent les 6400 neurones dits « d'arrivée » de la couche 1, leurs 100 core-tokens (tokens les plus activés en moyenne) respectifs, ainsi que les 10 neurones précurseurs (de la couche 0) présentant un poids de connexion le plus élevé associé à chaque neurone d'arrivée. Nous avons choisi de nous concentrer sur les 100 tokens les plus fortement activés en moyenne par chaque neurone, car il semble moins prioritaire, dans un premier temps, de s'intéresser aux tokens faiblement ou non activés par eux, étant alors en partie étrangers à l'extension de la catégorie associée à chaque neurone.

4.4 Objectif et mise en œuvre de l'étude en termes d'observables statistiques

La finalité de cette étude exploratoire est d'identifier des facteurs cognitifs synthétiques partiellement responsables de la segmentation catégorielle effectuée par les neurones formels. Ces facteurs sont mathématiquement intégrés dans la fonction d'agrégation neuronale et influencent l'identification des tokens constituant les core-tokens d'un neurone donné, autrement dit, la détermination des contenus de son extension catégorielle.

Plus en détail, nous visons à vérifier dans quelle mesure l'appartenance d'un core-token à la catégorie spécifique d'un neurone d'arrivée repose sur trois facteurs cognitifs que nous allons définir et proposer : l'amorçage, l'attention et le phasage catégoriels. Le niveau d'appartenance d'un core-token (dans la couche 1, soit la couche d'arrivée) à la catégorie associée à un neurone se mesurera à travers la valeur d'activation de ce token dans le neurone concerné. L'amorçage sera évalué à partir de la valeur d'activation d'un token dans ses neurones précurseurs respectifs (en couche 0). L'attention sera examinée via les valeurs des poids de connexion liant les neurones d'arrivée (couche 1) à leurs 10 principaux neurones précurseurs (ceux ayant les plus forts poids de connexion) dans la couche 0. Enfin, le phasage catégoriel sera quantifié en analysant le nombre de fois où un core-token au sein d'un neurone d'arrivée (couche 1) apparaît comme core-token parmi les 10 neurones précurseurs (couche 0) associés impliqués.

5 Définition des concepts cognitifs synthétiques étudiés et résultats

5.1 L'amorçage catégoriel synthétique

Dans le cadre de la psychologie humaine, l'amorçage [3, 18, 82, 38] est un processus cognitif via lequel un stimulus déclenche un premier niveau de réalisation

d'un processus cognitif, processus qui sera dès lors facilité, accéléré ou préparé dans le cadre de la réception d'un second stimulus, présentant un lien avec le premier. Plus spécifiquement, l'amorçage sémantique est le processus produisant le fait que la signification d'un élément (un mot par exemple) est rendue plus accessible à un individu par l'exposition préalable de celui-ci à un autre élément sémantiquement lié. L'effet d'amorçage est typiquement étudié en termes de délai de réponse dans le cadre de tâches de décision lexicale ou de compréhension de texte. Ce délai de réponse pouvant alors servir d'indicateur de l'existence, de la structure et de la force des relations sémantiques existant entre les mots et concepts de la mémoire sémantique à long terme.

La notion d'amorçage est liée à celle d'activation [47, 15, 46], postulant que des contenus ou processus cognitifs peuvent faire montre d'une intensité d'activité variable, avec des exemples prototypiques concernant des structures neuronales biologiques dont le niveau d'activité peut être physiologiquement « directement » mesurable (même si cette mesure est pour partie une reconstruction méthodologique et statistique). Dans le cas de l'amorçage, l'activation est pensée en termes de propagation de l'activation : une caractéristique cognitive (par exemple la signification) est « diffusée » d'une entité A (qui s'active en premier) à une entité B (qui s'active en résultante causale) (par exemple d'un mot à un autre) si A et B sont structurellement ou momentanément liés.

Nous postulons une transposition du concept d'amorçage, tel que défini ci-avant dans les champs des neurosciences et de la psychologie cognitive humaine, dans le domaine de la cognition synthétique. En effet, par construction mathématique de la fonction d'agrégation $\Sigma(w_{i,j}x_{i,j}) + a$ pour un élément donné (un token ou autre), la valeur d'activation de la catégorie portée par un neurone d'arrivée (sur une couche n) est *ipso facto* directement fonction (*modulo* la fonction d'activation) des valeurs d'activation $x_{i,j}$ des catégories associées à ses neurones précurseurs (sur la couche sous-ordonnée $n-1$). Autrement dit, en cohérence épistémologique avec la notion *princeps* d'amorçage, l'activation (lorsqu'elle existe pour un token donné) préalable des catégories vectorisées par des neurones précurseurs devrait « mathématiquement propager » l'activation de la catégorie associée à leur neurone d'arrivée correspondant. Nous formulons ainsi en ces termes une hypothèse d'amorçage catégoriel synthétique au sein des réseaux de neurones artificiels.

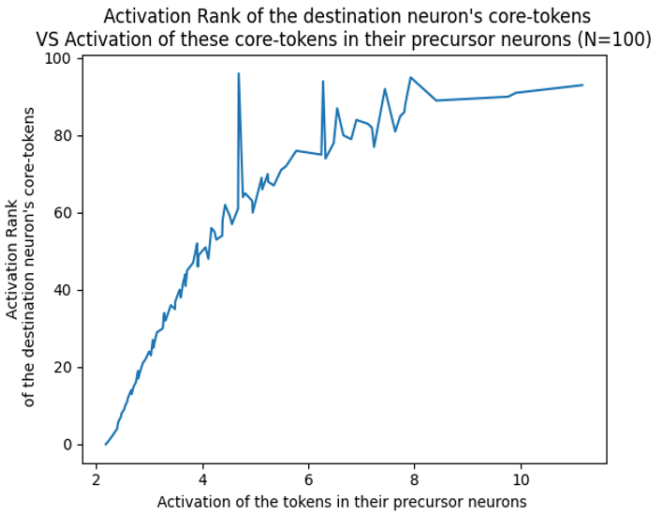
D'un point de vue quantitatif, l'observable empirique associée à notre hypothèse d'amorçage catégoriel synthétique est la valeur d'activation des neurones d'arrivée en fonction de leurs neurones précurseurs. Plus précisément, des données compatibles avec notre hypothèse devraient faire montre, pour une série de tokens donnée, d'une relation entre la valeur d'activation des neurones d'arrivée en couche $n+1$ et celle de leurs neurones précurseurs respectifs. Nous opérationnalisons cette démarche sur les 6400 neurones constitutifs de la couche 1 de GPT-2XL, en prenant en compte, pour chacun de ces neurones d'arrivée, de ses 10 neurones précurseurs à plus fort poids de connexion ainsi que de ses 100 tokens associés à une plus forte activation moyenne (core-tokens) (seuls les core-tokens activés dans au moins un précurseur sont pris en compte). D'un point de vue statistique, nous testons une relation de type ordinale (ρ de Spearman)entre

le rang moyen d'activation (variant de 1 à 100) des 100 core-tokens de chacun des 6400 neurones d'arrivée de la couche 1 et la moyenne des activations cumulées (i.e. additionnées) de ces tokens au sein des 10 neurones précurseurs qui leur sont associés (chaque core-token d'un neurone d'arrivée ayant une valeur nulle ou positive d'activation pour chacun des 10 précurseurs concernés).

Le tableau n°1 manifeste une relation positive, dotée d'une taille d'effet extrêmement forte (($\rho = .94$) et d'une significativité ($p < .001$). Le graphique n°1 illustre cette monotonie positive globale, avec cependant parfois des pics prononcés de variabilité. Le graphique n°2 en donne une vue pour un neurone d'exemple, avec une droite de régression montrant à nouveau une relation positive, bien que moins prononcée ici.

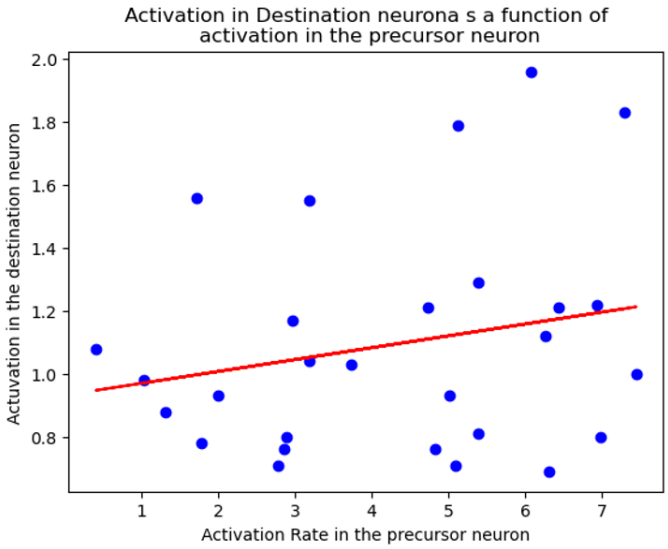
Strength of ρ	
ρ	.940
Significance of ρ	
p	.0000

Table 1 : Spearman's correlation between activation rank of the destination neuron's core-tokens and mean activation of these tokens in their precursor neurons (Layer 1, N_{max}=6400*100*10).



Graph n°1 : Activation rank of the destination neuron's core-tokens as a function of the mean activation of these tokens in their precursor neurons (Layer 1).

Ces résultats semblent compatibles avec notre hypothèse postulant un effet d'amorçage catégoriel synthétique, que nous nommons effet « x », à savoir une propagation mathématico-cognitive de l'activation des catégories neurorales précurseuses à leur catégorie neuronale associée d'arrivée en couche superordonnée.



Graph n°2 : Activation of the destination neuron's core-tokens as a function of the mean activation of these tokens in their precursor neurons (Layer 1, Control neuron 3000).

Ces cas d'illustration, à nouveau sans vocation de généralisation, nous permettent d'identifier comment le processus de phasage catégoriel permet d'extraire des catégories portées par les neurones précurseurs des sous-dimensions catégorielles co-activées qui vont dès lors constituer l'extension de core-tokens de leurs neurones d'arrivée corolaires.

5.2 L'attention catégorielle synthétique

En psychologie cognitive humaine, l'attention est définie comme un calibrage spécifique d'une activité en fonction de sa finalité, calibrage se traduisant par une plus grande efficacité des processus de prise d'information (dont la sélectivité) et d'exécution (dont la précision et la rapidité) à son endroit [59, 65, 58, 62, 68, 27, 67, 19, 81, 36]. Concernant la prise d'information externe, l'attention est liée à la conceptualisation [73, 74], c'est-à-dire à l'identification des seuls paramètres (des objets sur lesquels porte l'activité) dont la prise en compte est déterminante pour la réussite de l'activité ; paramètres sur lesquels l'action devra donc être calquée afin d'être ajustée et ainsi efficiente. L'attention est ici liée au filtrage et au formatage du nombre (trop) important d'informations perçues à disposition, c'est-à-dire à la non prise en compte (inhibition) de celles qui ne sont pas jugées comme pertinentes, afin de concentrer l'effort mental et la sélectivité informationnelle sur certains objets et propriétés. Concernant l'exécution de tâches, l'attention est liée au contrôle, par le système central, de l'activité consistant par exemple à attribuer un poids plus ou moins important (de priorité, d'ordre, de fiabilité, etc.) à certaines informations internes (con-

naissances, représentations, schémas) ou à vérifier la qualité de la réalisation de tâches dans le cadre de leur déroulement temporel.

D'un point de vue physiologique, l'attention est due aux capacités limitées de traitement de l'information du système nerveux et se traduit par des choix opérés dans l'intégration ou l'activation et l'exploitation de données sensorielles ou stockées en mémoire (sémantique, procédurale) [34, 6]. Ce qui va être réalisé par une réaction dite d'orientation consistant à concentrer les activités de recherche d'information vers un certain type de caractéristiques informationnelles.

Nous postulons ici une transposition du concept d'attention, comme présenté dans ce qui précède en psychologie cognitive et en neurosciences humaines, au domaine de la cognition artificielle. Cela dans la mesure où, par construction mathématique de la fonction d'agrégation $\Sigma(w_{i,j}x_{i,j})+a$, pour un élément donné (un token), sa valeur d'activation au sein de la catégorie associée à un neurone d'arrivée est *de facto* immédiatement dépendante (à la fonction d'activation près) des valeurs des poids de connexion $w_{i,j}$ de ce neurone d'arrivée avec ses neurones précurseurs. En d'autres termes, et en continuité épistémologique avec le concept originel d'attention, les poids de connexion avec les neurones précurseurs sont les régulateurs directs du niveau de prise de compte des informations (i.e. des niveaux d'activation) issues de ces neurones précurseurs, allant de l'inhibition ou du filtrage de ces données pour les poids négatifs, proches de 0 ou positifs mais faibles, à la focalisation et l'intégration mathématico-cognitive forte pour les poids importants. Dès lors, en termes maintenant d'exécution, les poids de connexion neuronaux pilotent le niveau d'exploitation qu'il est tenu pour pertinent par le système cognitif artificiel de réaliser des informations issues des catégories synthétiques antécédentes afin de réaliser la tâche courante d'un neurone successeur donné, tâche consistant à calculer le niveau d'appartenance d'un token courant à la catégorie constitutive de ce neurone superordonné. Nous délimitons dès lors en ces termes une hypothèse d'attention catégorielle synthétique à l'endroit des neurones artificiels, que nous dénotons sous le terme d'effet « w ».

5.2.1 Approche quantitative de l'attention catégorielle synthétique

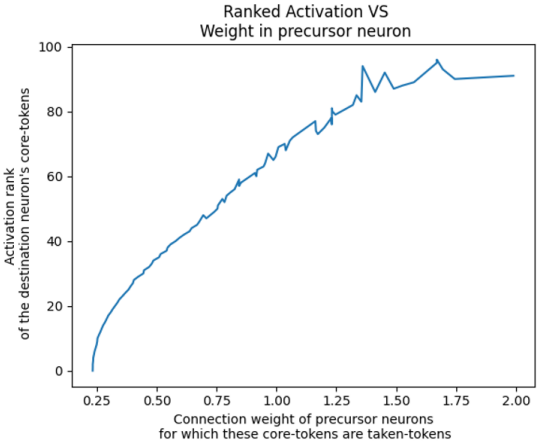
Au niveau quantitatif, l'observable empirique associée à notre hypothèse d'attention catégorielle synthétique est, pour un token donné, la valeur d'activation des neurones d'arrivée en fonction de leurs poids de connexion avec leurs neurones précurseurs respectifs. Une opérationnalisation de nature à tester cette hypothèse peut dès lors se fonder sur l'étude de la relation entre la valeur d'activation des neurones d'arrivée et les valeurs des poids de leurs connexions avec leurs neurones précurseurs. En effet, conformément à cette hypothèse, cette activation devrait croître avec l'augmentation des valeurs de ces poids antécédants. A nouveau, nous appliquons cette approche aux 6400 neurones constitutifs de la couche 1 de GPT-2XL, en prenant comme précédemment en compte, pour chacun de ces neurones d'arrivée, ses 10 neurones précurseurs à plus fort poids de connexion ainsi que ses 100 tokens associés à une plus forte activation moyenne (core-tokens) (précisons que seuls les core-tokens activés

dans au moins un précurseur sont exploités). D'un point de vue statistique, et de façon plus opérationnalisée, nous testons l'existence d'une relation ordinale (que nous allons mesurer avec un ρ de Spearman) entre (i) le rang moyen d'activation (variant de 1 à 100) des 100 core-tokens de chacun des 6400 neurones d'arrivée de la couche 1 et (ii) la moyenne des poids cumulés (i.e. additionnés) de connexion avec leur(s) (1 à 10) neurone(s) précurseur(s) respectif(s) pour lesquels ces tokens sont également des core-tokens.

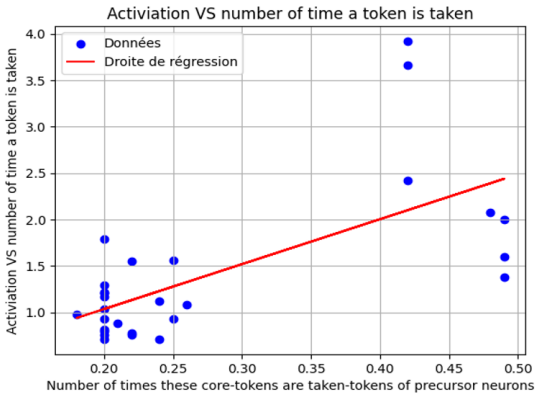
Nous pouvons observer sur tableau n°2 une relation ordinale positive entre le rang d'activation des neurones d'arrivée et la moyenne des poids de connexion cumulés avec leurs neurones précurseurs propres ; relation assortie d'une taille d'effet est extrêmement forte (ρ=.999) et d'une ample significativité (p<.001). Le graphique n°3 visualise cette relation monotone positive sur l'ensemble des données. Le graphique n°4 en fournit une exemplification pour un neurone témoin, avec une droite de régression montrant à nouveau une relation positive, bien que moins prononcée ici.

N_{max}	6400*100*10=6400000
n(mean ranks)	100
ρ	.999
p	.0000

Table 2: Spearman's correlation between activation rank of the destination neuron's core-tokens and mean connection weights with their precursor neurons (Layer 1).



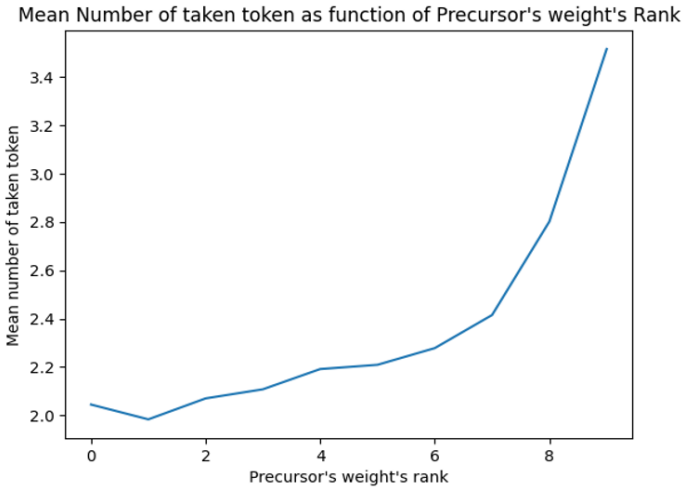
Graph n°3: Activation rank of the destination neuron's core-tokens as a function of the mean cumulative connection weights with their precursor neurons (Layer 1).



Graph n°4: Activation of the destination neuron's core-tokens as a function of the cumulative connection weights with their precursor neurons (Layer 1; control neuron 3000).

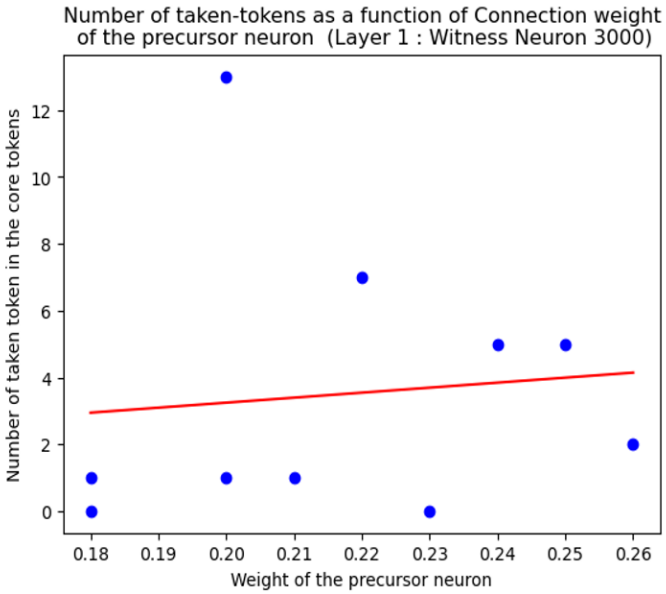
Les données exploratoires ici obtenues sont compatibles avec notre hypothèse d'attention catégorielle synthétique postulant une relation ordinale monotone positive entre niveau d'activation des core-tokens dans les neurones d'arrivée et poids de connexion de ces neurones d'arrivée avec ceux de leurs neurones précurseurs contenant ces mêmes tokens également comme core-tokens.

A nouveau dans le cadre d'une approche quantitative, tentons maintenant de comprendre plus avant la signification cognitive de cette attention catégorielle synthétique. Cela en s'interrogeant sur son *modus operandi* en termes de sélection d'information en entrée des neurones d'arrivée. Une question nous semblant extrêmement intéressante dans ce registre consiste à nous interroger sur la relation, pour un neurone d'arrivée donné, entre l'intensité de ses poids de connexion avec ses neurones précurseurs et le nombre de core-tokens « partagés » entre ce neurone d'arrivée et ces neurones précurseurs. Ce qui est équivalent à s'interroger de la sorte : dans quelle mesure les neurones précurseurs à forts poids de connexion alimentent-ils plus en core-tokens leurs neurones d'arrivée ; c'est-à-dire, dans quelle mesure les neurones précurseurs fortement connectés vont-ils plus présider à la constitution du contenu de l'extension (au sens catégoriel) de leurs neurones d'arrivée (i.e. à la constitution de leurs core-tokens) ; autrement dit encore, dans quelle proportion le poids de connexion régule la définition de l'extension et donc de la sélection et de la segmentation catégorielle opérée spécifiquement par un neurone (d'arrivée) synthétique donné. Nous obtenons en la matière une corrélation ordinale positive extrêmement forte ($\rho = .989, p < .001$) et significative (p<.001) entre d'une part (i) le rang moyen des poids de connexion de chaque neurone d'arrivée (du layer 1) avec ses neurones précurseurs et d'autre part (ii) le nombre moyen de core-tokens du neurone de destination qui étaient préalablement également des core-tokens des neurones précurseurs impliqués (cf graph n°6) ($n = 6,400$ neurones d'arrivée du layer 1 x 10 neurones précurseurs du layer 0 = 64000).



Graph n°5 : Mean number of taken-tokens as a function of percursor's weight's rank (Layer 1).

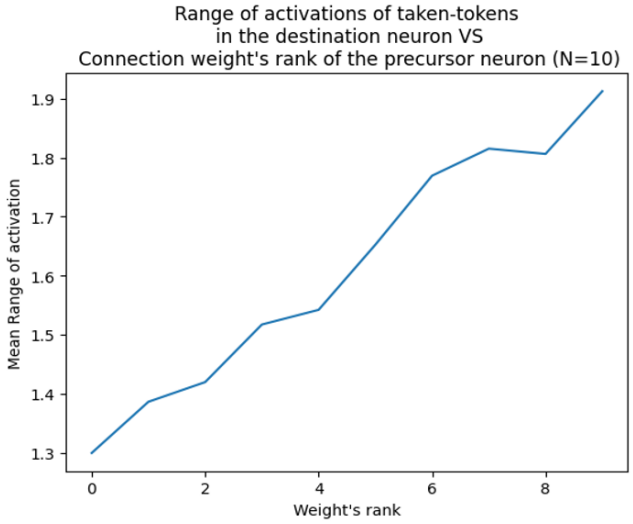
Nous nommons « taken-tokens » de tels tokens qui sont des core-tokens des neurones précurseurs et qui sont « repris » comme core-tokens par leurs neurones successeurs respectifs. Ce résultat, a nouveau assez cohérent étant donnée la nature de la fonction d'agrégation, nous indique que des poids attentionnels plus forts se traduisent par une sur-représentation de ces taken-tokens. Un poids attentionnel fort associé à un neurone précurseur tend ainsi à fonctionner, en terme de sélection de l'information, comme un « extracteur » d'une sous-dimension catégorielle (constituée des taken-tokens en jeu) à partir de la dimension catégorielle propre à ce neurone précurseur, sous-dimension qui va *de facto* génétiquement « alimenter » l'extension (de core-tokens) de la catégorie portée par le neurone d'arrivée concerné ; et ainsi participer à la segmentation catégorielle qui est spécifiquement la sienne. Le graph n°7 illustre cette tendance avec un neurone témoins du layer 1.



Graph n°6 : Number of taken-tokens as a function of percursor's weight (Layer 1, control neuron 3000).

Toujours d'un point de vue quantitatif, notons un autre résultat extrêmement intéressant et informatif afin de comprendre plus avant le mode de fonctionnement cognitif de l'attention catégorielle synthétique. De façon à nouveau cohérente avec la nature mathématique de la fonction d'agrégation, nous mesurons une corrélation ordinale positive, extrêmement forte ($\rho = .988, p < .001$) et significative (p<.001), entre d'une part le rang du poids de connexion précurseurs/successeurs et d'autre part l'étendue moyenne des activations dans les neurones d'arrivée des taken-tokens associés (ne sont pris en compte dans les calculs que les précurseurs associés à au moins 2 taken-tokens) (N_{max} = 6400 neurones d'arrivée du layer 1 x 100 tokens x 10 précurseurs). Tendance clairement illustrée par le graph n°8. Or, dans le champ de la catégorisation humaine, Thibault (1997) et Roads et al. (2024) indiquent, à propos du modèle de catégorisation « generalized context model » de Nosofsky (1986), que l'utilisation d'une distance pondérée (en l'occurrence de Minkowski) afin de rendre compte de la notion d'attention sélective est associée à une modification (contraction ou dilatation) de la métrique de l'espace de représentation catégorielle : des poids attentionnels faibles « rapprochent » les stimuli dans la dimension impliquée, alors que des poids élevés (attention forte) « étirent » l'espace de représentation catégorielle le long de cette dimension et augmentent ainsi la discrimination entre les stimuli impliqués. De façon transposée à la catégorisation synthétique, c'est bien ce que nous observons ici : un neurone d'arrivée ayant un fort poids de connexion avec un neurone précurseur donné voit ses taken-tokens issus de ce neurone précurseur avoir une variabilité d'activation plus impor-

tante dans l'espace d'activation de ce neurone d'arrivée ; autrement dit, les taken-tokens sont plus discriminés concernant leur niveau d'appartenance à la catégorie portée par ce neurone d'arrivée ; autrement dit encore, de forts poids de connexion augmentent l'empan de l'extension activationnelle des core-tokens des neurones d'arrivée en permettant de mieux différencier, de mieux contraster le degré d'appartenance d'un token donné à la catégorie impliquée. L'attention catégorielle synthétique serait donc associée au pouvoir discriminant d'une catégorie neuronale et, dès lors, à sa finesse analytique au sein de la dimension de segmentation de l'espace des tokens qui est la sienne. Cela, épistémologiquement en phase avec les caractéristiques conceptuelles mentionnées précédemment concernant la définition de l'attention en psychologie humaine.



Graph n°7 : Mean range of activation of taken-tokens in the destination neuron as a function of percursor's weight's rank (Layer 1).

Dans un registre quantitatif, les résultats empiriques de notre présente étude exploratoire tendent à être compatibles avec un phénomène de la cognition synthétique, celui de l'attention catégorielle artificielle, qualifié d'effet « w ». Cela, en pointant successivement trois caractéristiques potentielles de cette phénoménologie afférente à un poids de connexion attentionnel significatif : (i) la sélection, par un neurone d'arrivée, de certaines caractéristiques informationnelles spécifiques (i.e. certains types de core-tokens) (et pas d'autres) issues de ses neurones précurseurs, (ii) l'extraction associée, par un neurone d'arrivée, d'une sous-dimension particulière (de core-tokens) de la dimension catégorielle portée par chacun de ses neurones précurseurs et, (iii) le contraste permettant une différenciation plus fine (traduite par le niveau d'activation) de différents types d'éléments constitutifs de l'extension (de core-tokens) de la catégorie propre à un neurone d'arrivée.

5.2.2 Approche qualitative de l'attention catégorielle synthétique

Tentons maintenant de comprendre plus avant ces trois caractéristiques convergentes (qui ne sont *in fine* que des facettes alternatives l'une de l'autre) de l'attention catégorielle synthétique à partir d'une exploration qualitative de cette phénoménologie. Nous mobilisons à cette fin des exemples qualitatifs illustrant les modalités selon lesquelles les catégories portées par les neurones précurseurs associés à de forts poids attentionnels de connexion neuronale alimentent et génèrent sélectivement le contenu (en termes de core-tokens) de l'extension des catégories de leurs neurones d'arrivée respectifs. Cela, ainsi que nous allons l'observer, à travers un processus de « complémentation catégorielle » consistant à focaliser électivement l'attention du traitement calculatoire exécutoire de la fonction d'agrégation du neurone d'arrivée sur certaines sous-dimensions catégorielles extraites des neurones précurseurs afin de constituer la nature catégorielle propre à ce neurone d'arrivée, c'est-à-dire le contenu particulier de son extension catégorielle en termes de core-tokens.

Voici, à titre purement illustratif (cf tableau n°3), sans velléité d'exhaustivité, une comparaison de différents types catégoriels de core-tokens « apportés » sélectivement et respectivement, via un processus de complémentation catégorielle, par les différents neurones précurseurs à fort poids attentionnel de connexion, afin de constituer « progressivement », ajout sous-catégoriel par ajout sous-catégoriel, l'extension catégorielle propre à leur neurone d'arrivée associé. Nous identifions qualitativement deux principales classes de complémentation catégorielle : linguistique *versus* non linguistique.

Intéressons-nous en premier à la complémentation catégorielle de nature linguistique. Celle-ci peut être de nature sémantique, c'est-à-dire constituée d'ajouts catégoriels interprétables en termes d'opérations analogues à la sémantique humaine :

- **Par complémentation intra-lexicale**, consistant en une adjonction de tokens de même racine (variantes de tokenisation) ; exemple : un neurone précurseur « apporte » au neurone d'arrivée le token « manager », un autre le token « manag » et encore un autre le token « managerial ». Ou par complémentation intra-lexicale consistant en une adjonction de tokens issus de racines différentes ; exemple : un précurseur fournit au neurone d'arrivée le token « manager » et un autre le token « director » (le champ lexical est bien toujours le même ici).

- **Par complémentation sub-lexicale**, consistant en une adjonction de tokens issus d'un sous-champ lexical. Exemple : un précurseur alimente le neurone de destination avec le token «manager» et un autre avec les tokens «wenger», «Klopp» et «Mourinho» (ces derniers dénotent des entraîneurs de clubs de foot et relèvent bien à ce titre d'une sous-catégorie lexicale de «manager»).

- **Par complémentation péri-lexicale**, se traduisant par une adjonction de tokens en provenance d'un champ lexical connexe. Exemple :

un précurseur apporte «listen» alors qu'un autre fournit «sound»; ou un neurone précurseur délivre «order» et un autre «request»; ou encore un précurseur produit «necessary» quand un autre procure «indispensable».

- **Par complémentation para-lexicale**, via un ajout de tokens d'un champ lexical antonymique. Exemple : un neurone précurseur contribue au neurone d'arrivée avec «love», «adore» mais un autre avec «hate», «despise», «dislike».

La complémentation catégorielle linguistique peut également être de nature graphémique. Exemple : un neurone antécédant apporte au neurone d'arrivée le token « Said » et un autre antécédant le token « id » (nous retrouvons bien dans les deux cas le même graphème "id").

La complémentation catégorielle linguistique peut enfin être de type phonologique. Exemple : un précurseur amène les tokens « be », « bee » et un autre les tokens « Eve », « ea » (nous retrouvons bien dans les deux cas le même son /i/).

Penchons-nous maintenant sur la complémentation catégorielle non linguistique :

- **Celle-ci peut être de nature quantitative.** Exemple : un neurone précurseur alimente le neurone d'arrivée avec les tokens « er », « cv », « ku » et un autre avec les tokens « od », « fx », « yw » (chaque token contient invariablement exactement deux graphèmes).

- **Elle peut être de type culturel** (au sens d'éléments partagés par une culture humaine donnée). Exemple : un neurone apporte « ObamaCare » et un autre « Congres » (le Congrès des États-Unis ayant voté ce dispositif en mars 2010).

- **Elle peut également être d'autres natures**: non nécessairement interprétables par des catégories de pensée humaines mais liées à des contingences statistiques identifiées par le réseau neuronal durant son entraînement et donnant lieu à ce que nous dénoterons comme des « alien catégories » ou « non human like categories » ou encore des « catégories polysémiques » à partir de notre perspective cognitive humaine. Exemple : un neurone précurseur fournit le token « manager » alors qu'un autre associe le token « ID », sans que l'observateur humain puisse y assigner une logique (humaine) d'une nature ou d'une autre.

Ces exemples illustratifs, à nouveau sans volonté d'exhaustivité ou de systématicité, nous donnent à comprendre comment peut qualitativement être opéré le processus de sélection de l'information entrante par le mécanisme de l'attention catégorielle synthétique. Cela, à travers une activité de complémentation catégorielle permettant à la fonction d'agrégation d'un neurone d'arrivée d'extraire de chacun de ses neurones précurseurs à forts poids attentionnels de connexion une sous-dimension catégorielle particulière et contrastée par rapport

Complémentation catégorielle linguistique	Sémantique	Intra-lexicale	(manager) VS (manag) ; (manager) VS (managerial)
			(manager) VS (director)
		Sub-lexicale	(manager) VS (wenger, Klopp, Mourinho)
		Péri-lexicale	(listen) VS (sound) ; (order) VS (request) ; (necessary) VS (indispensable)
		Para-lexicale	(love, adore) VS (hate, despise, dislike)
	Graphémique		(Said) VS (ID)
	Phonologique		(be, bee) VS (Eve, ea)
Complémentation catégorielle non linguistique	Quantitative		(er, cv, ku) VS (od, fx, yw)
	Culturelle		(ObamaCare) VS (Congres)
	Autre		(manager) VS (ID)

Table n°3 : Exemples de modalités qualitatives de complémentation catégorielle (Layer 1).

aux autres ; l'apposition conjointe de ces sous-dimensions générant ainsi, sous-dimension catégorielle par sous-dimension catégorielle, le contenu catégoriel propre de l'extension du segment dimensionnel catégoriel que porte ce neurone d'arrivée.

5.3 Le phasage catégoriel synthétique

De par la construction de la fonction d'agrégation $\Sigma(w_{i,j}x_{i,j}) + a$nous postulons un troisième facteur mathématico-cognitif du niveau d'attribution d'un token à une dimension catégorielle neuronale. Facteur que nous nommons « phasage catégoriel synthétique », ou effet "Σ,", dans la mesure où la fonction d'agrégation d'un neurone d'arrivée additionne, pour un token donné, les valeurs pondérées de ses activations $w_{i,j}x_{i,j}$ au sein de ses neurones précurseurs respectifs. Plusieurs travaux en psychologie cognitive et neurosciences humaines impliquant la notion de phasage pourraient indirectement être ici l'objet d'analogies partielles avec la cognition synthétique dans ce registre ; par exemple, dans le champ des sujets de modalités perceptives [48, 41] ou de synchronisation cérébrale [1, 16, 61, 66, 64].

Dans le registre de la cognition synthétique qui est le nôtre, nous définissons la notion de phasage catégoriel synthétique par le fait qu'un token préalablement fortement activé pour différents neurones précurseurs (i.e. un core-token de ces neurones précurseurs) doit, à nouveau par construction mathématique de la fonction d'agrégation, être associé à un fort niveau d'activation au niveau du neurone d'arrivée associé ; cela, dans la mesure où ce token se retrouve *de facto* co-activé dans les différents termes constitutifs de la fonction d'agrégation ; co-activation dont la concaténation additive résulte en un niveau significatif d'activation de ce token en sortie du neurone de destination concerné. Un tel token est donc théoriquement l'objet d'un phasage des catégories neuronales des précurseurs en jeu : ces segments catégoriels précurseurs, bien que conceptuellement potentiellement différents, sont conjointement activés, ces dimensions rentrant alors en « écho » ou en « résonnance » catégorielle pour ce token précis ; cela, à travers une intersection catégorielle tracée au sein de ces dimensions et renforçant *ipso facto* le niveau d'activation en sortie dimensionnelle d'arrivée.

5.3.1 Approche quantitative du phasage catégoriel synthétique

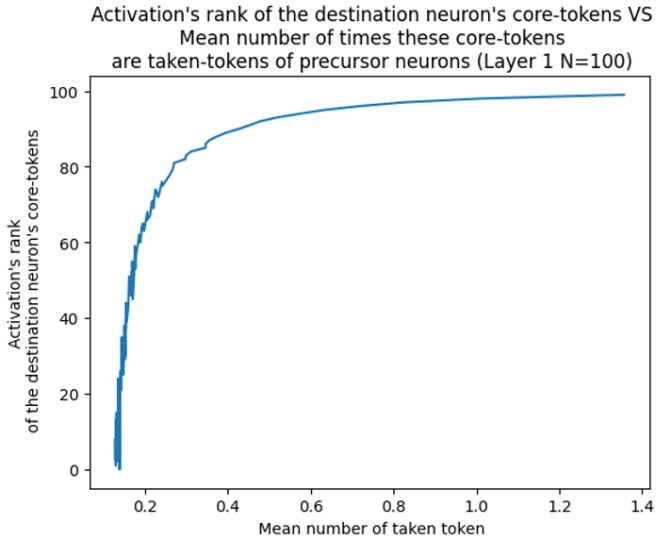
D'un point de vue quantitatif, nous opérationnalisons notre postulat de phasage catégoriel sous la forme de l'hypothèse suivante : plus un core-token d'un neurone d'arrivée est un core-token d'un nombre important de ses neurones précurseurs (à forts poids de connexion ici) et plus son niveau d'activation dans ce neurone d'arrivée est important. Hypothèse posant dès lors une relation monotone positive entre les deux variables impliquées. Le tableau n°4 présente les résultats compilés du test analytique de cette hypothèse à un niveau « local », c'est-à-dire au niveau de chacun des 6400 neurones d'arrivée isolément du layer 1. Nous y observons une corrélation ordinale entre les deux variables dotée d'une taille d'effet très forte (Mean $(\rho) = .976$), , largement significative (% of $(p(\rho) < .05) = 99.40\%$; $p(\chi^2) < .0001$)et massivement positive (% of $(\rho > 0) = 99.45\%$, $p(\chi^2) < .0001$). . Le tableau n°5, quant à lui, indique les résultats du test global de cette hypothèse directement au niveau de l'ensemble des données prises comme un tout (N_{max} =6400 neurones de la couche 1 x 10 précurseurs de la couche 0 x 100 core-tokens). Nous y trouvons à nouveau une corrélation ordinale entre les deux variables très forte ($\rho = .989$), positive et significative (p$p(\rho) < .001$). Le graph n°9 illustre graphiquement cette dernière tendance, en y manifestant une distribution logarithmique prononcée débouchant sur un plateau asymptotique ; et le graphe n°10 en fournit un exemple pour un neurone témoin assorti d'une droite de régression nettement positive.

Strength of ρ	
Mean (ρ)	.976
Significance of ρ	
% of (p(ρ)<.05)	99.40%
p(χ^2) of (p(ρ)<.05)	.0000
Positivity of ρ	
% of (ρ>0)	99.45%
p(χ2) of (ρ>0)	.0000

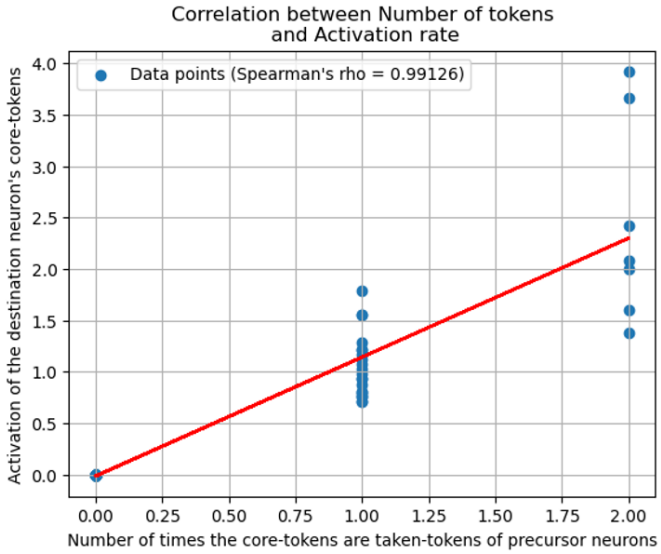
Table n°4 : Spearman's ρ correlation: Core-tokens' activation in the destination neuron VS Number of times these core-tokens are core-tokens of related precursor neurons (N=6400 ; Layer 1).

Strength of ρ	
ρ	0,989
Significance of ρ	
p(ρ)	0,000

Table n°5 : Spearman's ρ correlation: Core-tokens' activation's rank in the destination neurons VS Mean number of times these core-tokens are core-tokens of related precursor neurons (Layer 1).



Graph n°8 : Core-tokens' activation's rank in the destination neurons as a function of mean number of times these core-tokens are core-tokens of related precursor neurons (Layer 1).



Graph n°9 : Core-tokens' activation's rank in the destination neuron as a function of number of times these core-tokens are core-tokens of related precursor neurons (Layer 1, Control neuron 3000).

Dans une approche quantitative, l'ensemble des résultats ci-avant obtenu est compatible avec notre hypothèse de phasage catégoriel, qualifiée d'effet Σ,

postulant que plus un token est fortement activé (core-token) au niveau de plusieurs neurones précurseurs plus il tend à être fortement activé en sortie au niveau de leur neurone d'arrivée associé ; ce token se retrouvant dès lors à l'intersection catégorielle des catégories portées par ces précurseurs, catégories alors localement phasées.

5.3.2 Approche qualitative du phasage catégoriel synthétique

Tentons maintenant, au sein d'une dynamique cette fois-ci qualitative, de nous doter de points de repère de compréhension des modalités cognitives propres selon lesquelles des catégories, *a priori* distinctes ou tout du moins non isomorphes, associées à des neurones précurseurs peuvent se retrouver phasées catégoriellement localement, i.e. pour des tokens donnés. Cela afin de penser plus avant la phénoménologie conceptuelle à travers laquelle peuvent se manifester de telles intersections catégorielles, de tels croisements de segments catégoriels. Et donc de mieux nous représenter comment, via ces intersections, les neurones précurseurs alimentent et génèrent sélectivement l'extension des catégories de leurs neurones d'arrivée respectifs ; cela, en permettant l'extraction sélective de sous-dimensions catégorielles des catégories portées par les neurones précurseurs afin de constituer la nature catégorielle spécifique à leur neurone d'arrivée.

Dans une seule fin d'illustration, à nouveau sans volonté classificatoire systématique et encore moins exhaustive, le tableau n°6 présente des types d'exemples qualitatifs de modalités de phasage catégoriel. Cela, nécessairement, dans des cas où différentes catégories au niveau des neurones précurseurs sont conjointement activées pour des mêmes tokens donnés ; co-activations fortes provoquant génétiquement l'activation significative de la catégorie du neurone d'arrivé associé ou, autrement dit, co-activations définissant génétiquement le contenu (en termes de tokens) de l'extension catégorielle de cette catégorie d'arrivée (pour rappel, l'extension d'une catégorie étant définie ici, dans une perspective d'alpha-coupe en logique floue, par les 100 tokens les plus activés, à savoir les core-tokens).

Nous identifions qualitativement trois principaux types d'intersections catégorielles :

- **Le premier est de nature intra-lexicale** (relation d'identité sémantique). Exemple : deux catégories précurseures contiennent chacune, parmi leurs core-tokens respectifs, les mêmes tokens « manager » et « leadership », qui vont dès lors constituer une sous-dimension catégorielle extraite de la totalité de l'extension des deux dimensions catégorielles précurseures impliquées.

- **Le deuxième est de type sub-lexical** (relation d'inclusion sémantique). Exemple : une catégorie précurseure contient entre autres core-tokens « executive », « manager », « leader », « chief », « director », « CEO », « supervisor » ; et une autre : « director », « executive », « CEO ». Cette dernière série est incluse dans la première et va dès lors constituer

une sous-dimension catégorielle extraite des deux dimensions catégorielles précurseures.

- **La troisième est extra-lexicale**(relation de bi-lexicalité). Exemple : une catégorie précurseure contient dans son extension de core-tokens « knife », « gun », « mortar », « bomb », « axe », « cleaver », « sword », « grenade » (champ lexical des armes), et une autre : « cleaver », « spatula », « colander », « knife », « mixer », « mortar » (champ lexical des instruments de cuisine). A la croisée de ces deux champs lexicaux distincts, se trouvent les tokens « knife », « mortar » et « cleaver » qui vont dès lors constituer une sous-dimension catégorielle extraite de l'extension de core-tokens de ces deux dimensions catégorielles précurseures.

Intersection catégorielle intra-lexicale (identité)	(manager, leadership) VS (manager, leadership)
Intersection catégorielle sub-lexicale (inclusion)	(executive, manager, leader, chief, director, CEO, supervisor) VS (director, executive, CEO) (= top management)
Intersection catégorielle extra-lexicale (bi-lexicalité)	(knife, gun, mortar, bomb, axe, cleaver, sword, grenade) VS (cleaver, spatula, colander, knife, mixer, mortar)

Table n°6 : Exemples de modalités qualitatives de phasage catégoriel (Layer 1).

Ces cas d'illustration, à nouveau sans vocation de généralisation, nous permettent d'identifier comment le processus de phasage catégoriel permet d'extraire des catégories portées par les neurones précurseurs des sous-dimensions catégorielles co-activées qui vont dès lors constituer l'extension de core-tokens de leurs neurones d'arrivée corolaires.

5.4 Vision d'ensemble sur les trois facteurs de la segmentation catégorielle

Nous avons postulé l'existence de trois facteurs cognitifs synthétiques qui génèrent pour partie la segmentation catégorielle spécifiquement opérée par un neurone formel. Facteurs qui sont mathématiquement incarnés dans la fonction d'agrégation neuronale et qui, à ce titre, président (avec la fonction d'activation) à la détermination des seuls tokens qui constitueront les core-tokens d'un neurone donné, c'est-à-dire le contenu de son extension catégorielle. Ces trois facteurs que sont l'amorçage, l'attention et le phasage catégoriels pilotent ainsi le découpage catégoriel que les neurones vont opérer au sein de l'univers des tokens.

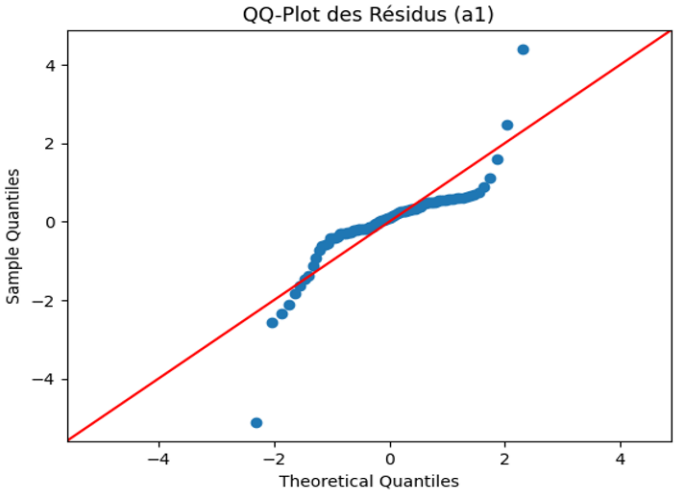
Afin de nous doter d'une sommaire représentation quantitative globale de l'action conjointe de ces trois facteurs, une régression linéaire multiple réalisée sur le rang d'activation des core-tokens dans les neurones d'arrivée en fonction (i) du nombre moyen de fois où ces core-tokens sont également des core-tokens des neurones précurseurs impliqués (a_1) (effet Σ), (ii) du poids moyen de connexion des neurones d'arrivée avec leurs neurones précurseurs associés (a_2) (effet w), et (iii) de l'activation moyenne de ces core-tokens dans les neurones précurseurs

concernés (a_3) (effet x)Précisons que cette régression, à des fins de faisabilité statistique, est réalisée uniquement sur les core-tokens des neurones d'arrivée qui sont core-token d'au moins un neurone parmi les neurones précurseurs en jeu ; et que lorsqu'un core-token d'arrivée est un core-token au sein de plusieurs précurseurs, son poids associé est la somme des poids précurseurs impliqués ; et qu'enfin, il en est de même concernant son activation au sein de ces précurseurs. Indiquons également que cette régression est réalisée sur les 6400 neurones constitutifs du layer 1.

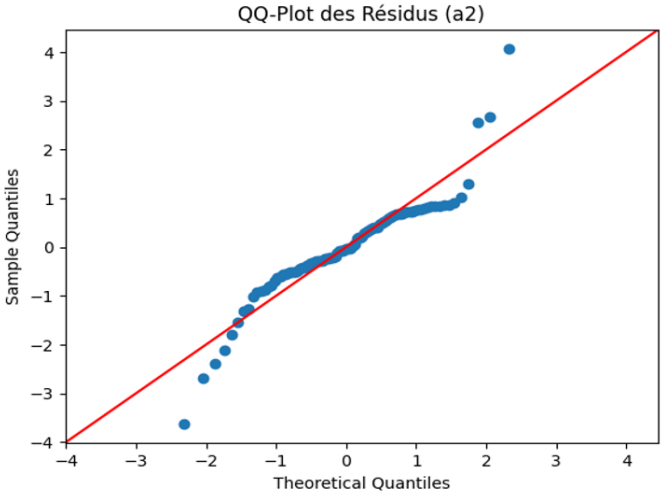
Cette régression linéaire (cf. tableau n°7) fait montre de coefficients linéaires standardisés positifs et non négligeables pour les trois facteurs postulés (s-a_1 = .86, s-a_2 = .56, s-a_3 = .65); cela, en phase avec nos hypothèses à leur endroit. Nous pouvons également noter que les impacts respectifs de ces trois variables indépendantes sur la variable dépendante semblent eux également non négligeables et de surcroit du même ordre de grandeur ($r^2(a_1) = .74$, $r^2(a_2) = .75$, $r^2(a_3) = .54$), nous donnant ainsi à penser que les trois facteurs identifiés impactent à travers des poids analogues la segmentation catégorielle opérée par les neurones d'arrivée. Même si ces présents résultats sont incertains dans la mesure où nos tests de normalité (Shapiro-Wilk, Kolmogorov-Smirnov et Jarque-Bera) des résidus ne sont pas en phase avec les conditions d'application attendues (ce que les graphs n°11 à n°13 semblent confirmer en mettant notamment à jour des outliers). Précisons de plus que nous suspectons ici des effets perturbateurs de co-linéarité entre les trois facteurs, ces derniers étant *a priori* fortement corrélés. Ces présents résultats sont donc uniquement à exploiter à titre illustratif.

Normality of Regression Residuals	
$p(SW_1)$	.000
$p(KS_1)$	.001
$p(JB_1)$	.000
$p(SW_2)$	.000
$p(KS_2)$	.067
$p(JB_2)$	.000
$p(SW_3)$	.000
$p(KS_3)$	.000
$p(JB_3)$	.000
Coefficients of the Linear Relashionship	
a_1	.271
standardized-a_1	.862
$r^2(a_1)$	.741
a_2	.555
standardized-a_2	.867
$r^2(a_2)$	.749
a_3	.648
standardized-a_3	.738
$r^2(a_3)$	.540

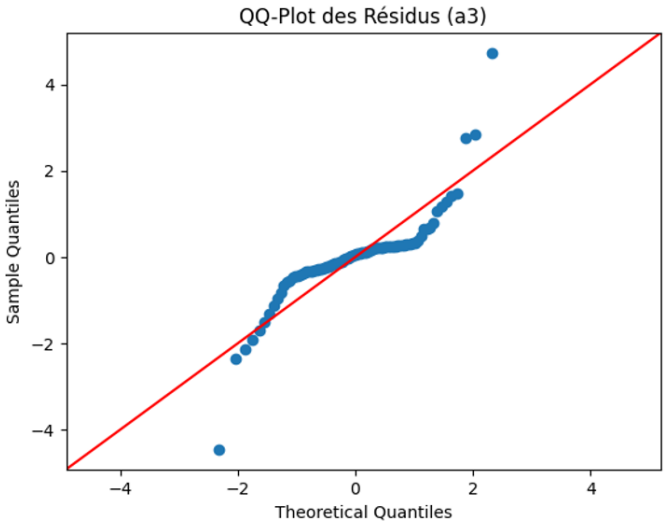
Table n°7 : Multiple linear regression of the activation rank of core-tokens in the target neurons based on (i) the average number of times they are core-tokens in the precursors, (ii) the average connection weight with the precursors, and (iii) the average activation in the precursors.



Graph n°10 : QQ-plot diagram of regression's residuals for a1 factor.



Graph n°11 : QQ-plot diagram of regression's residuals for a2 factor.



Graph n°12 : QQ-plot diagram of regression's residuals for a3 factor.

6 Conclusion

Dans le cadre de cette étude exploratoire, nous avons investigué, quantitativement mais aussi qualitativement, des facteurs génétiques de la segmentation catégorielle (du monde des tokens) opérée par les neurones synthétiques.
En nous basant sur la fonction d'agrégation $\Sigma(w_{i,j}x_{i,j}) + a$, nous avons postulé, par construction mathématique de celle-ci, trois facteurs mathématico-cognitifs impliqués dans ce découpage catégoriel. Le premier, l'amorçage catégoriel synthétique ou « effet x », est associé au fait que l'activation préalable des catégories vectorisées par des neurones précurseurs se propage à l'activation de la catégorie associée à leur neurone d'arrivé correspondant, impactant dès lors directement son extension catégorielle. Le deuxième, l'attention catégorielle synthétique ou « effet w », relève du fait que les poids de connexion entre un neurone d'arrivée et ses neurones précurseurs pilotent le niveau d'importance et d'exploitation alloué aux catégories précurseures dans la constitution de l'extension de la catégorie d'arrivée ; cela se traduisant, qualitativement, par un processus de complémentation catégorielle. Enfin, le phasage catégoriel synthétique ou effet Σ, ayant trait au fait que lorsque des segments catégoriels précurseurs, potentiellement conceptuellement différents, sont conjointement activés pour un token donné, ces dimensions rentrent alors en « écho » catégoriel en participant ainsi à la détermination du contenu de l'extension des catégories d'arrivée concernées ; cela se manifestant pour un processus d'intersection catégorielle.

Ces trois facteurs mathématico-cognitifs de la segmentation synthétique semblent génétiquement piloter un mécanisme d'extraction, à partir des catégories précurseures des neurones sous-ordonnés, de sous-dimensions catégorielles spécifiques ; ces dernières, qui combinées par la totalité de la fonction d'agrégation

(complétée par la fonction d'activation), façonnent le contenu (i.e. les core-tokens) de l'extension des catégories synthétiques ainsi résultantes au niveau des neurones hyper-ordonnés associés. Ce processus d'extraction conceptuelle synthétique, largement étudié en psychologie cognitive au niveau de son corollaire humain [13, 37, 31, 33, 9, 45, 84], se révèle passionnant d'un point de vue épistémologique et constitutif de la « construction du réel » opérée par la cognition synthétique, dans la fabrication qui est la sienne des arguments et prédicats du monde de tokens avec lequel elle interagit.
Nous étudions actuellement plus avant ce thème, dans une prochaine étude à paraitre, à travers l'investigation du processus d'abstraction catégorielle réalisé à partir des neurones précurseurs (couche n) par leurs neurones successeurs associés (couche n+1). Cela, en tentant de mieux comprendre comment un « détourage catégoriel », généré et piloté par les trois facteurs mathématico-cognitifs causaux que nous avons définis ici, est opéré sur la diversité catégorielle relative des core-tokens constitutifs de l'extension de la catégorie de chaque neurone précurseur afin d'extraire, de chacun d'entre eux, un sous-ensemble de tokens catégoriellement homogènes vis-à-vis de et alignés avec la catégorie spécifique que fabrique singulièrement leur neurone d'arrivée corollaire.

Remerciements

Les auteurs remercient Albert Yefimov (Sberbank & National University of Sciences & Technologies of Moscow) pour les stimulantes réflexions philosophiques et épistémologiques en matière d'IA avec lui. Les auteurs remercient également Madeleine Pichat pour sa relacture de cet article.

References

[1] Protachevicz, P. R., Hansen, M., Iarosz, K. C., Caldas, I. L., Batista, A. M., & Kurths, J. (2021). Emergence of neuronal synchronisation in coupled areas. *Frontiers in Computational Neuroscience*, 15, 663408. DOI: 10.3389/fncom.2021.663408.

[2] Schmalzried, M. (2024). The need of a self for self-driving cars: a theoretical model applying homeostasis to self driving. *arXiv preprint arXiv:2407.12795*. DOI: 10.48550/arXiv.2407.12795.

[3] Anderson, J. R. (1985). *Cognitive Psychology and Its Implications* (2nd ed.). W. H. Freeman. DOI: 10.4324/9781315784786

[4] Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7352–7364). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.656.

[5] Barkan, R. (2021). The Role of Cognitive Biases in Human Decision Making. *Journal of Behavioral Decision Making*, 34(3), 243–255. DOI: 10.1002/bdm.2210.

[6] Barr, W., & Bieliauskas, L. A. (2024). Neuropsychology of Decision Making: A Clinical Perspective. *Neuropsychology Review*, 34(1), 1–15. DOI: 10.1007/s11065-023-09500-1.

[7] Barsalou, L. W. (1995). *Cognitive Psychology: An Overview for Cognitive Scientists*. Lawrence Erlbaum Associates. DOI: 10.4324/9781315784786

[8] Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., & Filippova, K. (2022). "Will You Find These Shortcuts ? " A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.64

[9] Bathia, N., & Richie, D. (2024). Advances in Reinforcement Learning: Applications and Challenges. *Artificial Intelligence Review*, 57(2), 123–145. DOI: 10.1007/s10462-023-10123-4.

[10] Beaufils, M. (1996). Les réseaux de neurones artificiels: Modèles et applications. *Revue d'Intelligence Artificielle*, 10(4), 365–387. DOI: 10.1016/S0992-499X(97)80001-2.

[11] 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

[12] Bloch, H. (1992). *Grand dictionnaire de la psychologie.*

[13] Bolognesi, M. (2020). *Where Words Get Their Meaning: Cognitive Processing and Distributional Modelling of Word Meaning*. John Benjamins Publishing Company. DOI: 10.1075/ftl.7

[14] 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. Available from https://proceedings.mlr.press/v202/bricken23a.html

[15] Burns, R. B., & Graff, K. (2021). *Theories of Psychotherapy and Counseling: Concepts and Cases* (6th ed.). Pearson. DOI: 10.4324/9781315784786.

[16] Canales-Johnson, A., Silva, C., Huepe, D., Rivera-Rei, Á., Noreika, V., Del Carmen Garcia, M., Silva, W., Vaucheret, E., Sedeño, L., Couto, B., Melloni, M., Ibáñez, A., Chennu, S., Bekinschtein, T. A. (2015). Auditory feedback differentially modulates behavioral and neural markers of objective and subjective performance when tapping to your heartbeat. *Cerebral Cortex*, 25(11), 4490–4503. DOI: 10.1093/cercor/bhv076.

[17] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*. DOI: 10.48550/arXiv.2009.07896.

[18] Chao, L. L. (2024). Advances in Neuroimaging Techniques for Cognitive Neuroscience. *Journal of Cognitive Neuroscience*, 36(1), 1–15. DOI: 10.1162/jocn_a_01700.

[19] Cowan, N. (2024). Working Memory Capacity: Theories and Applications. *Annual Review of Psychology*, 75, 1–25. DOI: 10.1146/annurev-psych-010723-120001.

[20] Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., & Wei, F. (2022). Knowledge Neurons in Pretrained Transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. `https://doi.org/10.18653/v1/2022.acl-long.581`

[21] Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, D. A., & Glass, J. (2019, January). What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI, Oral presentation)*.

[22] Dalvi, F., Khan, A. R., Alam, F., Durrani, N., Xu, J., & Sajjad, H. (2022). Discovering Latent Concepts Learned in BERT. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2201.10020.

[23] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2010.00711`

[24] Dar, S. A., Durrani, N., Sajjad, H., Dalvi, F., & Belinkov, Y. (2023). Probing Pre-trained Language Models for Temporal Knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. DOI: 10.18653/v1/2023.acl-long.123.

[25] Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, (2019). Gradient descent finds global *minima* of deep neural networks, 1675-1685.

[26] 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

[27] Duncan, J. (1984). Selective Attention and the Organization of Visual Information. *Journal of Experimental Psychology: General*, 113(4), 501-517. DOI: 10.1037/0096-3445.113.4.501

[28] Durrani, N., Sajjad, H., Dalvi, F., & Belinkov, Y. (2022). On the Transformation of Latent Space in Fine-Tuned NLP Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.18653/v1/2022.emnlp-main.123.

[29] 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

[30] Enguehard, J. (2023). Extrmask: A Method for Explaining Time Series Predictions by Masking. *arXiv preprint arXiv:2301.08552*. DOI: 10.48550/arXiv.2301.08552.

[31] Eysenck, M. W., & Keane, M. T. (2020). *Cognitive Psychology: A Student's Handbook* (8th ed.). Psychology Press. DOI: 10.4324/9780429449229.

[32] 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

[33] National Centre for Nuclear Research. (2024). *41st International Free Electron Laser Conference (FEL2024)*. Warsaw, Poland. Retrieved from `https://fel2024.org/`

[34] Funayama, T., & Shibata, K. (2024). Advances in Quantum Computing: A Comprehensive Review. *Journal of Quantum Information Science*, 12(1), 45–67. DOI: 10.4236/jqis.2024.121004.

[35] Geva, M., Schuster, R., Berant, J., & Levy, O. (2023). Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.2012.14913.

[36] Gresch, D., & Müller, K. (2024). Machine Learning in Materials Science: Recent Progress and Emerging Applications. *Advanced Materials*, 36(5), 2105678. DOI: 10.1002/adma.202105678.

[37] Haslam, S. A., Reicher, S. D., & Platow, M. J. (2020). *The New Psychology of Leadership: Identity, Influence, and Power* (2nd ed.). Routledge. DOI: 10.4324/9781351108225.

[38] Hernández-Gutiérrez, C. A., & Pérez-González, J. (2024). Deep Learning Techniques for Natural Language Processing: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2), 1234–1256. DOI: 10.1109/TNNLS.2023.3101234.

[39] Howell, D. C. (2008). *Fundamental Statistics for the Behavioral Sciences* (6th ed.). Wadsworth Publishing. DOI: 10.1111/j.1467-985X.2008.00508_14.x.

[40] 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

[41] Capuano, F., & Kaup, B. (2024). Pragmatic Reasoning in GPT Models: Replication of a Subtle Negation Effect. Proceedings of the Annual Meeting of the Cognitive Science Society, 46. Retrieved from https://escholarship.org/uc/item/22q5920s

[42] 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`

[43] 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

[44] Ma, F., Plazyo, O., Billi, A. C., Tsoi, L. C., Xing, X., Wasikowski, R., Gharaee-Kermani, M., Hile, G., Jiang, Y., Harms, P. W., Xing, E., Kirma, J., Xi, J., Hsu, J., Sarkar, M. K., Chung, Y., Di Domizio, J., Gilliet, M., Ward, N. L., et al. (2023). Single cell and spatial sequencing define processes by which keratinocytes and fibroblasts amplify inflammatory responses in psoriasis. *Nature Communications*, 14(1). `https://doi.org/10.1038/s41467-023-39020-4`

[45] Marconato, E., & al. (2024). BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts. arXiv preprint arXiv:2402.12240. DOI: 10.48550/arXiv.2402.12240.

[46] Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2024). Implicature priming, salience, and context adaptation. *Cognition*, 244, 105667. DOI: 10.1016/j.cognition.2023.105667.

[47] Maxfield, M. G., & Babbie, E. R. (1997). *Research Methods for Criminal Justice and Criminology* (2nd ed.). Wadsworth Publishing. DOI: 10.4324/9781315784786

[48] Mitchell, M. (2021). *Abstraction and analogy-making in artificial intelligence. Annals of the New York Academy of Sciences*, 1505(1), 79-101. DOI: 10.1111/nyas.14619

[49] 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`

[50] Mousi, B., Durrani, N., & Dalvi, F. (2023). Can LLMs facilitate interpretation of pre-trained language models? *arXiv preprint arXiv:2305.13386*. DOI: 10.48550/arXiv.2305.13386.

[51] Nadeau, R. (1999). *Vocabulaire technique et analytique de l'épistémologie*. Presses universitaires de France.

[52] Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*. DOI: 10.48550/arXiv.2309.00941.

[53] Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.

[54] Paolo, G., Gonzalez-Billandon, J., & Kégl, B. (2024). A call for embodied AI. *arXiv preprint arXiv:2402.03824*. DOI: 10.48550/arXiv.2402.03824.

[55] 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. Available online: `https://www.youtube.com/watch?v=kG9Uv8-7OyQ&list=PLD25p-Bh6_swAk-TrFgk4lIQ6MQ2r5NTv&index=3`

[56] Pichat, M. (2024a). Psychologie de l'IA et alignement cognitif. Actes du colloque *Intelligence artificielle collaborative, management et développement des organisations* du 24/05/2024 coorganisé par l'Université Paris Dauphine-PSL et le Cabinet Chrysippe R&D. Available online: `https://www.youtube.com/watch?v=9TMmgbELaxQ&list=PLD25p-Bh6_sz6Sr7ms643GpCWW2LlIqeQ&index=6`

[57] Pichat, M. (2024). Psychology of Artificial Intelligence: Epistemological Markers of the Cognitive Analysis of Neural Networks. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2407.09563`

[58] Posner, M. I. (1978). *Chronometric Explorations of Mind*. Lawrence Erlbaum Associates.

[59] Posner, M. I., & Snyder, C. R. R. (1975). Attention and Cognitive Control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 55-85). Lawrence Erlbaum Associates. DOI: 10.4324/9781315784786

[60] Raieli, S., Altahhan, A., Jeanray, N., Gerart, S., & Vachenc, S. (2024). Escaping the Forest: Sparse Interpretable Neural Networks for Tabular Data. *arXiv preprint arXiv:2410.17758*. DOI: 10.48550/arXiv.2410.17758.

[61] Ribary, U., & Ward, L. M. (2024). Synchronization and functional connectivity dynamics across TC-CC-CT networks: Implications for clinical symptoms and consciousness. In *Phenomenological Neuropsychiatry: How Patient Experience Bridges the Clinic with Clinical Neuroscience* (pp. 105–118). Cham: Springer International Publishing. DOI: 10.1007/978-3-031-38391-5_10.

[62] Richard, J. C. (1980). *The Language Teaching Matrix*. Cambridge University Press.

[63] Roads, B. D., & Love, B. C. (2024). Modeling Similarity and Psychological Space. *Annual Review of Psychology*, 75(1), 215–240. DOI: 10.1146/annurev-psych-040323-115131.

[64] Rzechorzek, A. (2024). Understanding Cognitive Processes: Insights from Recent Research. *Journal of Cognitive Neuroscience*. DOI: 10.1162/jocn_a_01678.

[65] Schneider, W., & Shiffrin, R. M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention. *Psychological Review*, 84(1), 1-66.

[66] Shavikloo, M., Esmaeili, A., Valizadeh, A., & Madadi Asl, M. (2024). Synchronization of delayed coupled neurons with multiple synaptic connections. *Cognitive Neurodynamics*, 18(2), 631-643. DOI: 10.1007/s11571-023-10013-9.

[67] Tipper, S. P. (1985). The Negative Priming Effect: Inhibitory Priming by Ignored Objects. *The Quarterly Journal of Experimental Psychology*, 37A(4), 571-590. DOI: 10.1080/14640748508400920

[68] Treisman, A., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, 12(1), 97-136. DOI: 10.1016/0010-0285(80)90005-5

[69] Treviso, M., Lee, J., Ji, T., Van Aken, B., Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P. H., Martins, A. F. T., Forde, J. Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N.,. . . Schwartz, R. (2023). Efficient Methods for Natural Language Processing: A Survey. *Transactions Of The Association For Computational Linguistics*, 11, 826-860. `https://doi.org/10.1162/tacl_a_00577`

[70] Varela, F. (1984). The creative circle. In P. Watzlawick (Ed), *The invented reality*. London: W W Norton & Co Inc.

[71] Varela, F. J. (1988). *Cognitive Science: A Cartography of Current Ideas*. MIT Press.Varela1996

[72] Varela, F. J. (1996). Invitation aux sciences cognitives. Éditions du Seuil eBooks. http://inventin.lautre.net/livres/Varela-Invitation-aux-sciences-cognitives.pdf

[73] Vergnaud, G. (2009). Activité, développement, représentation. In M. Merri (Ed.), *Activité humaine et conceptualisation. Questions à Gérard Vergnaud* (pp. 149–154). Presses universitaires du Mirail.

[74] Vergnaud, G. (2016). Relations entre conceptualisations dans l'action et signifiants langagiers et symboliques. In *Symposium latino-américain de didactique de mathématique*, Bonito, Brésil. Disponible sur : https://www.gerard-vergnaud.org/texts/gvergnaud_2016_signifiants-langagiers-symboliques_conference-bonito.pdf.

[75] Voita, E., Sennrich, R., & Titov, I. (2021). Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. *arXiv preprint arXiv:2109.01396*. DOI: 10.48550/arXiv.2109.01396.

[76] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*. DOI: 10.48550/arXiv.2009.07896.

[77] Watzlawick, P. (1977). How real is real? London: Vintage Books.

[78] Watzlawick, P., Weakland, J. H., & Fisch, R. (1984). *Change: Principles of Problem Formation and Problem Resolution*. W. W. Norton & Company. DOI: 10.1002/9781119164894

[79] Wu et al., (2020). *pyOptSparse: A Python framework for large-scale constrained nonlinear optimization of sparse systems. Journal of Open Source Software*, 5(54), 2564. DOI: 10.21105/joss.02564

[80] Ji, M., & Wu, Z. (2022). *Automatic detection and severity analysis of grape black measles disease based on deep learning and fuzzy logic. Computers and Electronics in Agriculture*, 193, 106718.

[81] Wu, W. (2024). *We know what attention is!. Trends in Cognitive Sciences*, 28(4), 304-318.

[82] Xu, W., & Futrell, R. (2024). A hierarchical Bayesian model for syntactic priming. *arXiv preprint arXiv:2405.15964*. DOI: 10.48550/arXiv.2405.15964.

[83] Zadeh, L. A. (1996). Fuzzy Logic = Computing with Words. *IEEE Transactions on Fuzzy Systems*, 4(2), 103-111. DOI: 10.1109/91.493904

[84] Zettersten, M., Bredemann, C., Kaul, M., Ellis, K., Vlach, H. A., Kirkorian, H., & Lupyan, G. (2024). Nameability supports rule-based category learning in children and adults. *Child Development*, 95(2), 497-514. DOI: 10.1111/cdev.14008.

[85] Zheng, Y., & Stewart, N. (2024). Improving EFL students' cultural awareness: Reframing moral dilemmatic stories with ChatGPT. *Computers And Education Artificial Intelligence*, 6, 100223. `https://doi.org/10.1016/j.caeai.2024.100223`

[86] 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)*. DOI: 10.48550/arxiv.2309.01029.

[87] Zhang, Z., Song, Y., Yu, G., Han, X., Lin, Y., Xiao, C., ...& Sun, M. (2024). ReLU ^{2} Wins: Discovering Efficient Activation Functions for Sparse LLMs. *arXiv preprint arXiv:2402.03804*. DOI: 10.48550/arXiv.2402.03804.