Modélisation computationnelle de la familiarité absolue et relative selon un apprentissage de type Hebbien ou anti-Hebbien

Auteur : Warnier, William
Promoteur(s) : Sougné, Jacques; Delhaye, Emma
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en sciences psychologiques, à finalité spécialisée
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24733

Faculté de Psychologie, Logopédie et Sciences de l'éducation

MODÉLISATION COMPUTATIONNELLE DE LA FAMILIARITÉ ABSOLUE ET RELATIVE SELON UN APPRENTISSAGE DE TYPE HEBBIEN OU ANTI-HEBBIEN

Mémoire de fin d'études en vue de l'obtention du grade de Master en Sciences Psychologiques, à finalité spécialisée en Neurosciences cognitives et Psychologie Expérimentale

Promoteur : Sougné Jacques
Co-promotrice : Delhaye Emma
Superviseur : Read John

Année académique : 2024-2025

Remerciements

Je tiens à exprimer ma sincère gratitude à mes promoteurs pour leur accompagnement, leurs conseils avisés et leur disponibilité tout au long de ce travail. J'ai beaucoup apprécié nos longues discussions.

Je remercie également ma maman pour sa précieuse relecture et son soutien constant, qui m'ont grandement aidé dans la réalisation de ce mémoire.

Tables des matières

1 Introduction

1.1 La mémoire

La mémoire humaine est un système complexe et crucial pour le fonctionnement quotidien, permettant l'encodage, le stockage, et la récupération d'informations. Elle intervient dans pratiquement toutes les facettes de notre vie, de la reconnaissance simple des objets ou des visages jusqu'à la réalisation de tâches complexes impliquant un raisonnement, une planification ou la prise de décision. Grâce à ce processus cognitif essentiel, les individus sont en mesure de retenir et d'utiliser des expériences passées pour interagir efficacement avec leur environnement présent et anticiper les événements futurs (Tulving, 1985). Elle est traditionnellement divisée en deux grandes catégories : la mémoire explicite (ou déclarative) et la mémoire implicite (ou non-déclarative) (Squire, 2004).

1.1.1 La mémoire explicite

La mémoire explicite inclut des souvenirs qui sont conscients et qui peuvent être verbalisés, y compris des faits et des informations (Eichenbaum & Cohen, 2001). Elle est elle-même divisée en mémoire épisodique et mémoire sémantique (Tulving, 1972). La mémoire épisodique permet de se souvenir des évènements spécifiques de sa propre vie (e.g. votre premier jour à l'Université), tandis que la mémoire sémantique concerne les connaissances générales sur le monde (e.g. Phnom Penh est la capitale du Cambodge).

1.1.2 La mémoire implicite

La mémoire implicite concerne quant à elle les habiletés et les réponses acquises à travers l'expérience, telles que rouler à vélo ou lacer ses chaussures, lesquelles sont réalisées sans réflexion consciente (Milner et al., 1998). Cette mémoire inclut aussi les habitudes et le conditionnement par lequel les réponses émotionnelles sont apprises et déclenchées inconsciemment.

1.2 La mémoire de reconnaissance

La mémoire de reconnaissance, appartenant à la mémoire déclarative, nous permet de déterminer si nous avons déjà fait l'expérience d'un évènement, ou d'un objet (Yonelinas, 2024). Par exemple, identifier sa voiture parmi les autres sur un parking ou reconnaitre la maison de notre enfance sur une photo. Ce travail se concentrera sur la reconnaissance visuelle et n'abordera pas les reconnaissances auditive, olfactive ou encore sensitive.

Il est courant de distinguer la reconnaissance selon deux catégories : la reconnaissance avec souvenir conscient d'éléments de contexte et la reconnaissance sans souvenir d'éléments de contexte, ces deux catégories ont été respectivement nommées recollection et familiarité (Yonelinas, 1994, 2002; Yonelinas et al., 2024). Un exemple régulièrement cité dans la littérature pour illustrer cette distinction est celui du « boucher dans le bus » (Mandler, 1980), dont voici une déclinaison : imaginez-vous assis dans une salle de cinéma. Une personne monte les marches de l'allée et vient se placer sur le siège juste devant vous. Vous ressentez un sentiment particulier, vous savez que vous avez déjà vu cette personne mais vous ne pouvez pas retrouver son nom ou toutes autres informations à son sujet. Cette première reconnaissance, automatique, s'apparente au sentiment de familiarité. Vous fouillez ensuite votre mémoire afin de déterminer qui est cette personne ainsi que l'endroit où vous l'avez vue. Enfin, cela vous revient, cette personne est le capitaine de l'équipe de football que vous avez affrontée ce week-end. Cette seconde reconnaissance, tenant compte quant à elle du contexte qui entoure la personne ou l'objet reconnu et ayant fait appel à des processus stratégiques de recherche en mémoire, est appelée recollection.

Cette distinction entre familiarité et recollection n'a toutefois pas toujours fait consensus dans la littérature. En effet, la reconnaissance était d'abord décrite comme sous-tendue par un seul processus (Egan, 1958). Depuis lors, plusieurs auteurs s'accordent pour affirmer que deux processus sont inclus dans la reconnaissance : la recollection et la familiarité (Eichenbaum et al., 2007). Certains auteurs (Jacoby, 1991; Juola et al., 1971; Mandler, 1980; Tulving, 1985; Yonelinas, 1994) ont par ailleurs proposé des modèles de la mémoire de reconnaissance avec divers niveaux de complexité et parfois quelques différences, notamment concernant la vitesse ou encore la synchronicité des processus. Par exemple, dans leur modèle, Atkinson et Juola (1974) suggèrent que la familiarité a lieu avant la recollection et donc que les processus s'exécutent de façon sérielle, tandis que d'autres (Jacoby, 1991; Mandler, 1980; Yonelinas,

1994) proposent que les deux processus sont initiés en parallèle, mais leur différence de vitesse de résolution amène à des différences quant aux vitesses de réponse.

La théorie en deux processus de la reconnaissance s'appuie sur plusieurs arguments. Premièrement, des études montrent que la familiarité fonctionne plus rapidement que la recollection, ce qui suggère un fonctionnement différentiel. En effet, lorsque les temps-limite de réponse sont courts, les sujets peuvent faire des discriminations basées sur la familiarité plus vite que celles nécessitant la recollection d'informations spécifiques (Hintzman & Caulton, 1997 ; Gronlund, Edwards & Ohrt, 1997).

Deuxièmement, la recollection et la familiarité présentent des corrélats électrophysiologiques distincts. L'électro-encéphalographie (EEG) est une méthode utilisée en neuro-imagerie pour étudier les potentiels évoqués (ERP, *Event-Related Potentials*) et qui permet d'examiner l'activité électrique du cerveau en réponse à des stimuli ou des évènements. Elle offre une résolution temporelle très précise, de l'ordre de la milliseconde. Des études utilisant des ERP indiquent que les items « rappelés », associés à une mémoire précise de certains détails de l'évènement, montrent des distributions temporelles et spatiales sur le scalp qui sont différentes des ERP liées aux items reconnus sur base de la familiarité (Curran, 2000 ; Düzel et al.,1997). Plus précisément, la recollection est associée à une composante pariétale positive située entre 400 et 800 ms après la présentation du stimulus. En revanche, la familiarité est caractérisée par une composante frontale négative apparaissant plus précocement, entre 300 et 500 ms. Ces différences temporelles et topographiques confirment que la recollection et la familiarité reposent sur des processus neurocognitifs distincts (Curran, 2000).

Troisièmement, les données issues de la neuropsychologie clinique soutiennent l'existence d'une double dissociation entre recollection et familiarité sur le plan neuroanatomique. Plusieurs études ont montré que certaines lésions cérébrales altèrent sélectivement l'un ou l'autre de ces deux processus. Par exemple, des lésions touchant le fornix, un faisceau majeur reliant l'hippocampe au corps mamillaires et au cortex préfrontal, sont associées à une altération marquée de la recollection, tandis que la familiarité demeure préservée (Aggleton et al., 2000 ; Tsivilis et al., 2008). A l'inverse, , la patiente NB, une épileptique ayant subi une ablation chirurgicale de la partie antérieure du lobe temporal gauche incluant le PrC, avec hippocampe préservé, montre une atteinte spécifique de la familiarité, en l'absence de déficit significatif de recollection (Bowles et al., 2007 ; Köhler & Martin, 2020). Ces observations soutiennent un modèle dans lequel la recollection dépend principalement de

l'hippocampe et de ses connexions, tandis que la familiarité s'appuie sur des régions parahippocampiques antérieures, en particulier le PrC (Aggleteon & Brown, 1999 ; Yonelinas et al., 2002). Ce schéma lésionnel croisé fournit un appui fort en faveur de la dissociation fonctionnelle et neuroanatomique entre ces deux composantes de la mémoire de reconnaissance.

1.2.1 La recollection

La recollection est un processus qui inclut la récupération d'éléments du contexte associé à un événement (Yonelinas, 2002; Diana et al., 2007). La recollection fonctionnerait selon un processus binaire. Un souvenir est soit recollecté avec des détails, soit il ne l'est pas, auquel cas la reconnaissance doit s'appuyer sur la familiarité. La recollection contribue donc de manière qualitative à la reconnaissance en rappelant des aspects spécifiques de l'expérience. Pour évaluer la recollection, plusieurs techniques ont été utilisées. Dans la procédure *Remember/Know/Guess* (Gardiner, 1988), les participants indiquent si la reconnaissance est accompagnée de la récupération de détails contextuels (*Remember*, « je me souviens avoir vu cet item ») ou pas (*Know*, « je sais que j'ai vu cet item »). La réponse *Guess* permet au sujet d'identifier les réponses données en devinant et ainsi minimiser le risque de réponses *Know* basées sur l'incertitude et non pas sur la familiarité. Les réponses *Remember* seraient donc un indice de recollection. Dans un article d'opinion, Diana et al. (2007) présentent une corrélation positive entre l'intensité du signal BOLD (*blood oxygen level dependent*) dans l'hippocampe avec les réponses de type recollection. Cela suggère que plus un stimulus a de chance d'être associé à une réponse *Remember* ou d'être récupéré avec des détails contextuels, plus l'activité dans l'hippocampe augmente. D'autre part, le paradigme de mémoire source offre une méthode objective pour évaluer la recollection en mémoire de reconnaissance (Besson et al., 2012). Il consiste à présenter des items dans des contextes d'encodage distincts, puis à tester non seulement la reconnaissance de l'item mais aussi la récupération du contexte associé. Une réponse correcte à la question sur la source implique que le participant a accédé à une trace mnésique riche et spécifique de l'épisode original, ce qui correspond au processus de recollection. Ce paradigme est particulièrement utile pour distinguer les réponses fondées sur un sentiment de familiarité de celles basées sur une recollection.

1.2.1 La familiarité

Selon Yonelinas (2002), la familiarité est conceptualisée comme un processus qui permet aux individus de reconnaitre un stimulus comme déjà vu ou connu et qui ne nécessite pas l'accès conscient à des détails de contexte de l'expérience originale. Contrairement à la recollection, la familiarité ne fonctionnerait pas de façon binaire mais plutôt de façon quantitative avec un gradient de détection de signal (Yonelinas, 2010). Ainsi, pour chaque item présenté lors de la phase de reconnaissance d'un test de mémoire, le sentiment de familiarité pour cet item sera plus ou moins fort selon que l'item ait été étudié ou non. Les jugements de familiarité sont rapides et automatiques, sans engagement de ressources cognitives importantes (Yonelinas, 2002).

La contribution de la familiarité dans la reconnaissance est évaluée notamment grâce aux paradigmes à temps de réponse limité. Ils consistent à imposer un délai très court pour la réponse au moment de la reconnaissance d'un stimulus. L'objectif est de contraindre les participants à s'appuyer uniquement sur les processus les plus rapides, en l'occurrence la familiarité, tout en empêchant ou limitant le recours à la recollection. Dans le paradigme Remember/Know/Guess présenté plus haut, les réponses Know sont considérées comme indicateurs directs de la familiarité. Elles sont particulièrement sensibles aux manipulations expérimentales qui influencent la familiarité, comme la répétition des items. (Yonelinas et al., 2010).

Pour juger de leur familiarité, les stimuli visuels sont d'abord traités par la voie visuelle ventrale. La voie visuelle ventrale, aussi appelée voie du « quoi », s'étend des aires visuelles primaires vers le cortex temporal inférieur (voir figure 1).

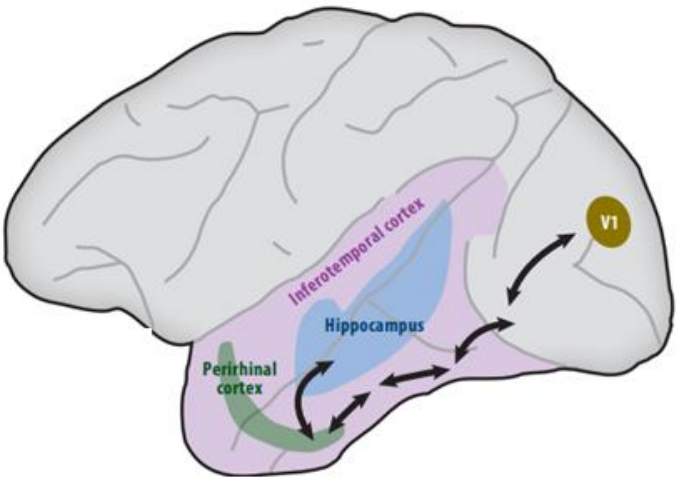

Figure 1. Voie visuelle ventrale au départ des aires visuelles primaires (V1) passant par le cortex temporal inférieur jusqu'au PrC.

Tout au long de cette voie ventrale, les représentations des stimuli se complexifient, passant de lignes et d'angles vers des formes combinées, et la signification sémantique se précise (Murray & Bussey, 1999). Cette complexification s'accompagne d'une augmentation du champ de réponse des cellules. Le champ visuel pour lequel chaque cellule s'active est réduit au début de la voie et s'étend en avançant vers le pôle temporal. Cet agrandissement de la zone de perception des cellules permet de complexifier le traitement et d'appréhender des objets complexes en entier pour en déduire la signification (Kravitz et al., 2013). A l'arrivée de cette voie ventrale se situe le PrC, où la trace mnésique précédemment stockée sera mise en relation avec la représentation du stimulus présenté (Suzuki & Naya, 2014). Ainsi, d'un point de vue anatomique, la structure la plus souvent mise en lien avec la familiarité est le PrC.

Le PrC, situé à l'interface entre le cortex temporal médian et la voie visuelle ventrale, possède des connexions bidirectionnelles avec ces deux systèmes ainsi qu'avec des régions corticales associatives (Eichenbaum et al., 2007; Suzuki & Naya, 2014). Son implication dans la reconnaissance visuelle des objets est largement documentée (Brown & Aggleton, 2001; Murray & Richmond, 2001). Dans le modèle intégratif de la mémoire proposé par Bastin et al. (2019), le PrC participe à la représentation percepto-conceptuelle d'objets uniques et complexes, et son activation soutient le sentiment de familiarité. Ce rôle repose sur sa capacité à coder des entités, c'est-à-dire des exemplaires uniques d'une catégorie, définis par la conjonction spécifique de leurs attributs perceptifs et conceptuels. Le PrC permettrait ainsi une forme de séparation de patterns appliquée aux objets (Gardette et al., 2025), essentielle pour discriminer finement des stimuli très similaires, indépendamment de leurs variations visuelles (e.g., orientation, luminosité). Il s'agit donc d'une structure-clé dans la représentation de ces entités jouant un rôle au sein d'un système plus large impliqué dans le traitement, la liaison et la récupération des attributs spécifiques des souvenirs (Bastin et al., 2019). De plus, le PrC joue un rôle fondamental concernant la fluence — définie comme la facilité avec laquelle un stimulus est traité. Cette fluence, accrue pour les stimuli déjà rencontrés, peut concerner le traitement perceptuel ou conceptuel du stimulus (Mandler, 1980), et elle est souvent utilisée comme signal heuristique pour juger de la familiarité. La proposition récente de Gardette et al. (2025) renforce cette conception en suggérant que le PrC est activé lors de jugements de familiarité liés à des entités, et que cette activation reflète une combinaison entre force représentationnelle et fluence. En somme, le PrC apparaît comme un nœud fonctionnel essentiel

où s'articulent représentation d'entités complexes et émergence de la fluence, contribuant au sentiment de familiarité.

1.2.2 Tests de reconnaissance

Plusieurs tâches testant la reconnaissance ont été utilisées dans la littérature du domaine (Besson et al., 2012). Chaque test de reconnaissance commence par une phase d'encodage, durant laquelle des stimuli sont présentés et appris par les participants. Dans ce type de tâches, les sujets mémorisent les stimuli présentés dans la phase d'apprentissage de façon explicite ou implicite (conscient vs. inconscient), c'est-à-dire avec la consigne claire qu'ils doivent étudier les stimuli ou non. A la suite de l'encodage, les participants réalisent une phase de reconnaissance. Si la façon dont cette seconde phase diffère selon le paradigme choisi (oui/non, choix forcés, etc.), le principe sous-jacent reste identique. Plus précisément, des anciens stimuli (i.e., des stimuli familiers) ainsi que de nouveaux stimuli sont présentés successivement aux participants, lesquels doivent déterminer pour chaque stimulus s'il a déjà été vu lors de la phase d'apprentissage.

La reconnaissance oui/non se base sur un paradigme dans lequel le sujet affirme ou non reconnaitre le stimulus présenté, après avoir vu précédemment un ensemble de stimuli durant une phase d'apprentissage. Cette technique ne permet cependant pas de distinguer la contribution des différents processus de reconnaissance. Pour cela, la procédure *Remember/Know/Guess* est recommandée. Comme développé plus haut, elle permet de déterminer la contribution de chaque processus (familiarité et recollection) à la reconnaissance.

Une autre technique est la reconnaissance à choix-forcés durant laquelle le sujet se voit présenter au moins deux stimuli et doit choisir le stimulus déjà vu parmi le ou les distracteurs non-vus.

1.2.3 Théorie de détection du signal (SDT, *Signal Detection Theory*)

La SDT permet de modéliser les réponses binaires (« déjà vu » / « nouveau ») d'un système confronté à une incertitude perceptive (e.g. tâche de reconnaissance oui/non). Elle repose sur l'idée que les stimuli génèrent des activations, neuronales ou computationnelles, en fonction du système étudié, plus ou moins fortes selon qu'ils ont été précédemment rencontrés (« signal présent ») ou non (« signal absent »). Ces activations sont représentées par deux distributions de scores potentiellement superposées. La reconnaissance par familiarité ne repose pas sur un processus tout ou rien, mais sur le dépassement d'un seuil décisionnel appliqué à un

signal continu : si le signal excède ce seuil, l'image est classée comme «déjà vue» ; sinon, elle est classée comme «nouvelle». Ainsi, la décision de reconnaissance découle d'un traitement quantitatif et graduel du signal de familiarité. Une différence importante à remarquer est celle entre la force du signal, qui est une variable continue, et le seuil de décision, qui permet une catégorisation binaire basée sur le signal continu. Le seuil peut être interprété comme la force de signal minimum nécessaire pour attribuer le stimulus associé à la catégorie « déjà vu ». La force du signal est comparable au sentiment de familiarité ressenti, alors que le seuil reflète la décision comportementale. Dans la SDT, ce seuil est nommé Criterion noté c. Cette approche basée sur un seuil permet de distinguer quatre types de réponses représentés sur la figure 2. Les *Hits* sont la reconnaissance correcte d'un stimulus, les *Miss* sont la non-reconnaissance d'un stimulus étudié, les *False Alarms* (FA) sont les essais pour lesquels un distracteur est erronément reconnu comme un stimulus étudié et les *Correct Rejection*s (CR) sont le rejet correct d'un distracteur.

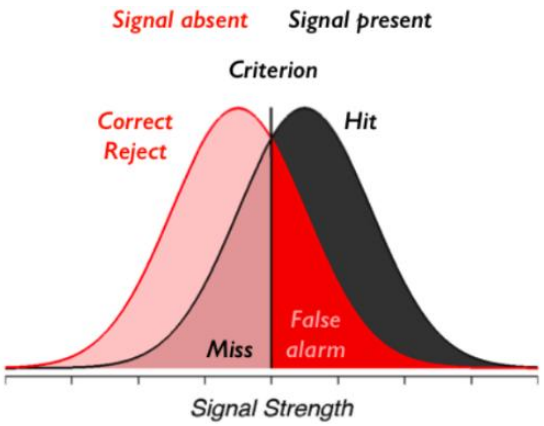

Figure 2. Illustration des quatre catégories de réponses sous forme d'aire sous la courbe, respectivement à droite et à gauche du Criterion/seuil.

Cette modélisation permet de quantifier la sensibilité du système à discriminer les signaux à l'aide de l'indice d', calculé selon la formule suivante :

$$d' = Z(\text{pHit}) - Z(\text{pFA})$$

où Z désigne la transformation inverse d'une distribution normale cumulée. Elle est particulièrement adaptée à l'analyse des performances dans des tâches de reconnaissance oui/non, comme celles simulées dans ce travail. La SDT permet le calcul d'un autre indice, le Criterion selon cette formule :

$$c = -\frac{1}{2}[Z(\text{Hit Rate}) + Z(\text{FA Rate})]$$

Chez l'humain, cet indice représente la tendance du répondant ainsi que le seuil de réponse. Si c = 0, le répondant est neutre dans ses décisions. Un c > 0 indique un biais conservateur, l'individu exige un signal de familiarité élevé pour répondre « déjà vu ». Un c < 0 indique une biais libéral, l'individu répond « déjà vu » pour un signal de familiarité plus faible. Dans le contexte de ce travail, qui se concentre sur l'expérimentation artificielle, les décisions ne sont pas comportementales mais directement inférées sur base du signal de familiarité, le modèle artificiel n'affiche donc pas de biais de réponse. Cela implique un c neutre (= 0) en toutes conditions. Cependant, si on utilise les valeurs brutes des signaux de familiarité, il est possible d'interpréter le c comme la valeur seuil déterminée implicitement par le modèle artificiel. Ces valeurs sont donc potentiellement comparables à travers différentes conditions. Pour illustrer, un modèle sera dit plus libéral en condition A par rapport à B, si le seuil en A se situe à 18000 et celui en B à 22000. Dans la littérature concernant la SDT, le d' et le c sont présentés comme indépendants, une modification de la discriminabilité n'impacte pas le criterion, ni inversement (Macmillan & Creelman, 2004). Dans ce travail, le criterion sera envisagé seulement suivant sa fonction de seuil car le biais de réponse sera toujours nul, c'est pourquoi le terme « seuil » sera utilisé pour désigner cet indice.

1.3 Familiarité absolue et relative

En 1980, Mandler a proposé une distinction conceptuelle majeure dans les mécanismes de la reconnaissance en mémoire : celle entre familiarité absolue et familiarité relative. Cette distinction émerge notamment de l'étude du *word frequency mirror effect* (WFME), selon lequel les mots de faible fréquence, rarement utilisés dans la langue, produisent davantage de *Hits* et moins de *FA* tandis que les mots de haute fréquence produisent plus de *FA* et moins de *Hits* (Joordens & Hockley, 2000; Reder et al., 2000). Ceci s'observe dans une tâche de jugement « nouveau/déjà vu » de mots parmi lesquels certains ont été vus lors de la phase d'encodage. Cet effet serait expliqué par deux phénomènes sous-jacents : (1) par des différences de discriminabilité (d'), avec un d' supérieur en condition de faible fréquence mais aussi (2) par un déplacement du criterion. En condition de haute fréquence, le criterion serait stratégiquement

élevé (plus conservateur), entrainant moins de *Hits* et plus de *FA*. A l'inverse, en condition de basse fréquence, le criterion est plus libéral, ce qui implique plus de *Hits* et moins de *FA* (Joordens & Hockley, 2000).

Mandler (1980) avance que ce phénomène s'explique par l'existence de deux types de familiarité : la familiarité absolue, accumulée au cours des multiples expositions à un item tout au long de la vie, et la familiarité relative, liée à une modification récente du niveau de familiarité d'un item suite à une exposition ponctuelle (Coane et al., 2011). Le WFME expose des performances différentes selon la fréquence des stimuli. Cette fréquence est comparable au niveau de familiarité absolue.

La familiarité absolue est un signal global, stable et durable, comparable à un compteur mnésique général des expositions passées. Elle est fortement corrélée à la fréquence lexicale et repose sur l'intégration d'attributs perceptifs et sémantiques sur le long terme (Mecklinger & Bader, 2020). A l'inverse, la familiarité relative correspond à l'élévation temporaire du signal de familiarité provoquée par une exposition récente (Coane et al., 2011). Prenons par exemple le mot « galéjade » qui désigne une histoire plaisante et exagérée. Son niveau de familiarité absolue est faible, il a rarement été traité par notre système cérébral. Par contre, suite à la lecture de ce mot dans ce texte, son niveau de familiarité relative vient d'augmenter fortement. Ce mécanisme a été expliqué par Coane et al. (2011) (voir figure 3). Après une exposition, le niveau de familiarité augmenterait selon deux composantes. Une augmentation forte mais temporaire représente la familiarité relative. La familiarité absolue montre une augmentation plus faible et plus durable après chaque exposition.

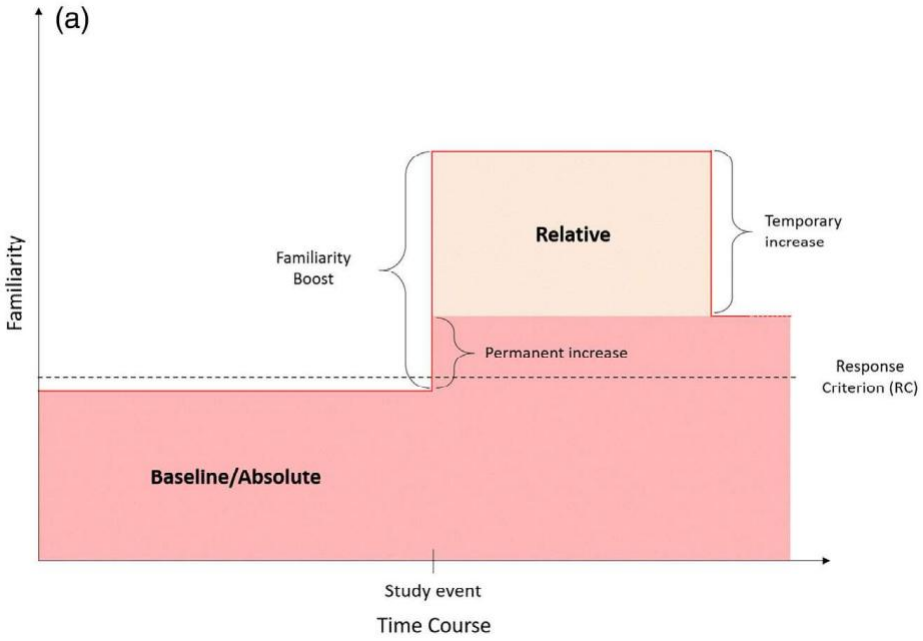

Figure 3. Evolution hypothétique du signal de familiarité suite à une exposition et contribution de la familiarité absolue (rouge) et relative (orange) au signal de familiarité total. Le seuil de décision (ici : « Response Criterion ») indiqué en pointillé. Reproduit de Read, J., Delhaye, E., & Sougné, J. (2024). Computational models can distinguish the contribution from different mechanisms to familiarity recognition. *Hippocampus*, *34*(1), 36-50.

La distinction entre familiarité absolue et relative se base également sur des études en électrophysiologie, notamment les ERP. Dans un article de revue, Mecklinger et Bader (2020) mettent en évidence que ces deux formes de familiarité sont associées à des composantes ERP distinctes. La familiarité absolue est souvent reflétée par une modulation de l'onde N400, une composante centro-pariétale de valence négative survenant environ 400 ms après la présentation du stimulus. Son amplitude est réduite pour les mots à haute fréquence lexicale (à haute familiarité absolue) en comparaison avec les mots rares. Cette réduction est interprétée comme un indice de fluence conceptuelle (Yonelinas, 2002), traduisant un traitement plus efficient des items connus.

En revanche, la familiarité relative est davantage associée à une modulation de l'onde FN400, une composante fronto-centrale négative survenant également autour de 300-500 ms après présentation du stimulus. Cette onde est sensible aux effets de récence et reflète une fluence de traitement inattendue : lorsqu'un stimulus récemment rencontré est traité plus facilement que prévu, l'activité FN400 est réduite. Ce mécanisme de surprise de fluence traduit

une mise à jour implicite du système mnésique, où la facilité de traitement perçue dépasse les attentes basées sur la familiarité absolue seule.

Ainsi, les deux formes de familiarité sont dissociables au plan temporel (voir figure 4) et topographique : la N400 indexe une familiarité basée sur les connaissances accumulées, tandis que la FN400 reflète une augmentation transitoire du sentiment de familiarité, propre à la familiarité relative.

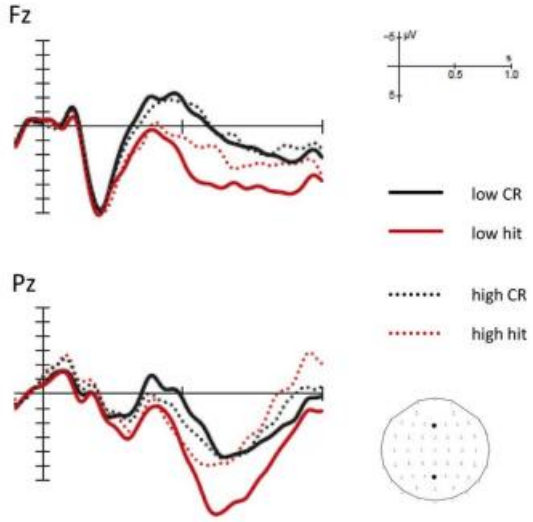

Figure 4. Composantes ERP frontale (Fz) et pariétale (Pz) montrant les Hits (en rouge) et les CR en noir, pour des items de basse fréquence (ligne continue) ou de haute fréquence (ligne pointillée) Reproduit de Mecklinger, A., & Bader, R. (2020). From fluency to recognition decisions : A broader view of familiarity-based remembering. *Neuropsychologia*, *146*, 107527.

Les données issues des neurosciences permettent aujourd'hui de distinguer la familiarité relative de la familiarité absolue à travers des marqueurs neuraux spécifiques, notamment dans le PrC. Par exemple, Yang et al. (2023) ont montré, par des analyses multivariées en IRMf, que le PrC encode distinctement et automatiquement les deux types de familiarité, celle issue d'une exposition récente (familiarité relative) et celle accumulée au cours de la vie (familiarité absolue), même lorsque ces informations ne sont pas pertinentes pour la tâche. Dans cette étude, les auteurs ont manipulé la familiarité relative en variant le nombre de répétitions de mots présentés dans une tâche incidente (jugement d'animalité, « oui/non »), tandis que la familiarité absolue a été évaluée par des jugements subjectifs portant sur la familiarité cumulative à long terme avec les concepts représentés par ces mots. Ces deux signaux coexistent mais s'expriment de façon dissociable au niveau du signal BOLD, ce qui suggère des mécanismes de traitement distincts au sein d'un même substrat neuroanatomique. . Duke et al. (2017) rapportent une

augmentation du signal BOLD dans le PrC avec l'augmentation de la familiarité absolue (évaluée par la fréquence subjective du mot), alors que Yang et al. (2023) montrent une diminution du signal pour des items récemment vus, ce qui renforce l'idée de processus différenciés mais complémentaires.

Ces résultats chez l'humain rejoignent les observations faites chez le primate non-humain. Dans une étude de référence, Xiang et Brown (1998) ont enregistré l'activité de neurones individuels dans la région temporale antérieure de macaques exécutant une tâche de reconnaissance visuelle. Ils ont identifié trois types distincts de neurones dans le PrC. Les neurones de nouveauté montrent une activité maximale lors de la première présentation d'un stimulus. Les neurones de récence ont une activité inversement proportionnelle au délai écoulé depuis la dernière présentation du stimulus et constituent donc un candidat plausible pour coder la familiarité relative. Les neurones de familiarité encodent la familiarité au long terme indépendamment de la récence. Ces derniers montrent une diminution progressive de leur réponse au fil des répétitions, correspondant à un signal lentement dégressif typique de la familiarité absolue. Cette différenciation neuronale entre récence et familiarité cumulative est essentielle pour comprendre comment le système mnésique est capable de distinguer un stimulus « déjà vu » récemment d'un stimulus familier en général, notamment par des mécanismes clairement distincts selon le type de familiarité.

Enfin, des divergences subsistent dans la littérature quant à la nature du codage de la familiarité dans le PrC. Gimbel et al. (2017) suggèrent un codage continu, avec des variations proportionnelles à la force du signal mnésique, tandis que Martin et al. (2016) défendent un codage catégoriel distribué, selon lequel le PrC représenterait la familiarité à travers des patterns activés de manière qualitative selon les niveaux de familiarité.

Le Prc est une région du cerveau précocement impactée dans la maladie d'Alzheimer (MA) (Juottonen et al., 1998). Les individus atteints de troubles cognitif léger amnésique sont potentiellement à un stade précoce de la MA et ont probablement des atteintes significatives au PrC. Le statut de la familiarité dans le trouble cognitif léger amnésique a fait l'objet de résultats contrastés dans la littérature, certaines études concluant à une préservation de la familiarité (Anderson et al., 2008 ; Besson et al., 2015), tandis que d'autres observent au contraire une altération de ce processus (Pitarque et al., 2016 ; Wolk et al., 2008). Cette divergence pourrait s'expliquer par les paradigmes utilisés : sans isoler clairement la contribution de la familiarité (par rapport à la recollection), un déficit spécifique peut passer inaperçu. De plus, une fois la

familiarité isolée par des tâches adaptées, il est important de dissocier les types de familiarité étudiées (absolue et relative). Des tâches ont donc été développées pour distinguer la familiarité relative, basée sur le nombre d'expositions récentes d'un stimulus, de la familiarité absolue, correspondant à l'impression de connaître un stimulus grâce à l'apprentissage accumulé au cours de la vie (Anderson et al., 2021). Par exemple, on peut demander aux participants d'estimer la fréquence de présentation récente d'un item (afin de mesurer la familiarité induite par répétition), ou de juger le degré de familiarité de concepts connus indépendamment du contexte d'apprentissage (afin d'évaluer la familiarité sémantique durable). Grâce à de telles méthodes, il a été montré que les patients aMCI présentent un déficit sélectif de la familiarité relative : ils sont significativement moins sensibles que des témoins aux augmentations de familiarité induites par des expositions récentes répétées (Anderson et al., 2021). En revanche, leur capacité à évaluer la familiarité absolue de concepts courants demeure globalement comparable à celle de sujets âgés en bonne santé.

Ce profil différentiel est cohérent avec des observations en neuroimagerie chez les aMCI, qui montrent une réduction de la suppression de répétition, c'est-à-dire de la diminution d'activité neuronale normalement observée lorsqu'un même stimulus est présenté plusieurs fois (Yu et al., 2016). Or, ce mécanisme est considéré comme un marqueur du codage neuronal de la familiarité récente dans le PrC ; sa perturbation suggère donc une altération spécifique de la familiarité relative dans cette population.

1.4 Modèles computationnels

Les modèles computationnels jouent un rôle essentiel en neurosciences cognitives, en fournissant un cadre formel pour explorer les mécanismes sous-jacents aux fonctions cérébrales. Dans le domaine de la mémoire (Kazanovich & Borisyuk, 2021; Norman & O'Reilly, 2003; O'Reilly et al., 2014), ils permettent de simuler les processus d'encodage, de consolidation et de récupération de l'information en s'appuyant sur des principes issus des neurosciences biologiques. Ces modèles offrent la possibilité de tester des hypothèses mécanistiques de manière explicite, de générer des prédictions vérifiables, et de relier les niveaux d'analyses comportemental, neuronal et synaptique. Ils contribuent ainsi à une compréhension intégrée du fonctionnement mnésique.

1.4.1 Apport des modèles computationnels

Depuis le Perceptron de Rosenblatt, qui effectuait des tâches de classifications binaires (Rosenblatt, 1958), plusieurs modèles artificiels ont tenté de reproduire le fonctionnement de systèmes biologiques. Ces modèles présentent plusieurs avantages pour faire avancer les neurosciences cognitives. Ils permettent de formaliser les théories cognitives en clarifiant les hypothèses puis en confrontant ces hypothèses aux données produites par ces modèles. Aussi, les modèles permettent d'articuler plusieurs niveaux de complexité, par exemple la biologie neuronale avec des comportements complexes comme la reconnaissance.

1.4.2 Historique des modèles de reconnaissance

Ces dernières décennies, plusieurs auteurs ont tiré profit de la modélisation pour faire avancer le domaine de la mémoire et de la reconnaissance. En 2003, Norman et O'Reilly ont proposé un modèle à deux versants sur base du constat que l'hippocampe serait une structure adaptée à la mémorisation rapide des épisodes de sorte à ce qu'ils puissent être rappelés ultérieurement sur base d'indices partiels tandis que le cortex temporal médian – dont fait partie le PrC – apprendrait lentement, par incrémentation, les régularités présentes dans son environnement. Nous ne nous intéresserons ici qu'à la partie du modèle consacrée au cortex temporal médian, c'est-à-dire le modèle néocortical.

Le modèle néocortical (Norman & O'Reilly, 2003) est un modèle computationnel à 2 couches (couche d'entrée et couche cachée) qui peut reconnaître les items étudiés des non-étudiés suivant leur représentation dans la couche cachée qui symbolise le PrC. Par une propagation vers l'avant (*i.e., feedforward)* de la couche d'entrée vers la couche cachée, le modèle permet de raffiner la représentation d'un stimulus dans le PrC au fur et à mesure des présentations. Ce processus d'affinage repose sur le fonctionnement de la loi Hebbienne (Hebb, 1949) :

« Quand un axone d'une cellule A est assez proche pour exciter une cellule B de manière répétée et persistante, une croissance ou des changements métaboliques prennent place dans l'une ou les deux cellules ce qui entraîne une augmentation de l'efficacité de A comme cellule stimulant B. ».

Le modèle néocortical de Norman et O'Reilly (2003) repose sur une règle d'apprentissage Hebbienne, selon laquelle les connexions synaptiques entre neurones co-activés sont renforcées. Ainsi, les neurones activés lors de la première présentation d'un

stimulus seront plus susceptibles de l'être à nouveau lors de sa réexposition. Par ailleurs, le modèle intègre un mécanisme de compétition locale via la règle du k-Winners-Take-All (kWTA), qui permet de contraindre l'activité dans la couche cachée en ne conservant que les les k neurones plus activés (où k est un pourcentage, par exemple, 10 % d'entre eux), les autres étant inhibés par rétroaction. Ce double mécanisme d'apprentissage synaptique et de régulation compétitive conduit à un raffinement progressif de la représentation des stimuli, en favorisant l'activation sélective d'un sous-ensemble stable de neurones. Lors du test, le score de familiarité est déterminé en calculant la moyenne de l'activation des k neurones les plus fortement activés dans la couche cachée. Cette approche reflète un signal continu de familiarité fondé sur la correspondance globale entre le stimulus testé et les représentations stockées. Ce comportement du modèle est compatible avec les propriétés observées chez certains neurones du cortex périrhinal, appelés neurones de nouveauté, qui montrent une forte réponse lors de la première exposition à un stimulus, suivie d'une diminution progressive de leur activité lors des présentations ultérieures (Brown & Xiang, 1998 ; Duke et al., 2017). Le modèle néocortical est donc particulièrement adapté pour capturer des phénomènes tels que l'augmentation cumulative du sentiment de familiarité avec l'exposition répétée, particulièrement par le parallèle évident entre l'apprentissage incrémentiel du réseau et l'augmentation graduelle de la familiarité absolue au cours des expositions répétées à un stimulus. Ce réseau basé sur un mécanisme Hebbien semble donc rendre compte de la familiarité absolue, mais sans encoder le processus moins cumulatif qu'est la familiarité relative.

Dans l'objectif de modéliser les neurones de nouveautés du PrC, Bogacz et Brown (2003a) ont proposé un modèle de neurones artificiels à apprentissage anti-Hebbien. Cette règle d'apprentissage implique une diminution des poids synaptiques en réponse à la co-activation d'un neurones d'entrée et d'un neurone de sortie, ce qui conduit à une réduction de l'activation des neurones de sorties lors de la seconde présentation d'un même stimulus. Dans leurs simulations, chaque stimulus visuel est représenté par un vecteur binaire sur un ensemble de neurones d'entrées. Les auteurs montrent que ce type d'apprentissage permet au réseau de reconnaitre jusqu'à 1200 stimuli différents avec une précision de 99%, même lorsque les patterns d'entrée sont fortement corrélés. Cette performance supérieure à celle des réseaux Hebbiens s'explique par le fait que le réseau anti-Hebbien encode préférentiellement les caractéristiques distinctives des stimuli, tandis que les modèles Hebbiens tendent à renforcer les caractéristiques communes à plusieurs stimuli (Bogacz & Brown, 2003a).

Dans un autre article de 2003 (Bogacz & Brown, 2003b), les mêmes auteurs comparent plusieurs types de réseaux modélisant la familiarité dans le PrC. Leur objectif est d'évaluer l'efficacité de différentes règles d'apprentissage dans des réseaux à une seule couche, soumis à des séquences d'entrées binaires simulant des stimuli sensoriels. Dans ces simulations, chaque réseau encode un ensemble de vecteurs binaires, représentant des patterns perceptifs, et doit ensuite signaler si un nouveau vecteur présenté est familier ou non. Les performances des modèles sont évaluées en fonction du taux d'erreurs de reconnaissance (*FA* et *Miss*) et de la capacité de mémoire, c'est-à-dire le nombre maximum de patterns pouvant être stockés et reconnus avec une erreur acceptable. Les auteurs constatent que les modèles anti-Hebbiens surpassent les modèles Hebbiens sur ces deux dimensions. Plus précisément, les réseaux anti-Hebbiens sont moins sensibles à l'interférence entre patterns similaires, ce qui leur permet de conserver une capacité de discrimination plus stable à mesure que le nombre d'items mémorisés augmente. Ces résultats soutiennent l'hypothèse que le PrC pourrait utiliser des mécanismes anti-Hebbiens pour la discrimination de la familiarité. De plus, Tyulmankov et al. (2022) ont implémenté un réseau qui utilise du méta-apprentissage pour déterminer quelle règle privilégier, Hebbienne ou anti-Hebbienne. Autrement dit, ce réseau apprend comment apprendre à détecter si un stimulus est familier ou non. Cet apprentissage est continu, sans phase explicite de test. Les résultats montrent que la règle anti-Hebbienne est préférentiellement choisie. Ce constat appuie la pertinence de l'anti-Hebbien malgré sa plausibilité biologique qui n'a pas encore été démontrée.

1.4.3 Modèle ResHebb

Plus récemment, Read et al. (2024) ont implémenté un modèle s'inspirant des travaux de Kazanovich & Borisyuk (2021), combinant deux étapes distinctes lors de la présentation d'un stimulus pour apprentissage : le réseau convolutif ResNet50 (He et al., 2016) pré-entrainé, imitant les processus observés au cours de la voie visuelle ventrale (LeCun et al., 2015), suivi d'un réseau à propagation vers l'avant (i.e., *feedforward*) à deux couches, nommé module de mémoire. C'est dans ce module que l'apprentissage a lieu lors des simulations. Ce modèle a été utilisé pour comparer les règles d'apprentissage Hebbienne et anti-Hebbienne sur des images réelles. L'architecture de ce modèle est reprise pour le travail présent et illustrée dans la figure 5. La description technique du modèle se trouve dans la partie méthode.

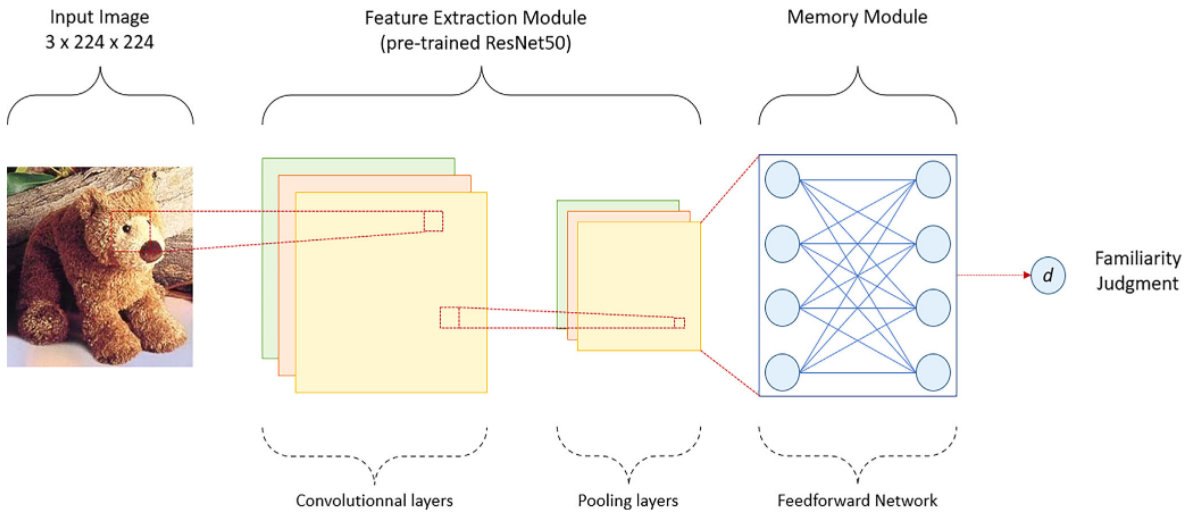

Figure 5. Architecture du modèle de Read et al. (2024). Reproduit de Read, J., Delhaye, E., & Sougné, J. (2024). Computational models can distinguish the contribution from different mechanisms to familiarity recognition. *Hippocampus*, *34*(1), 36-50.

Avant de procéder aux tests de jugement de familiarité, les réseaux sont entrainés durant une phase d'apprentissage. C'est-à-dire qu'ils modifient leurs poids entre les couches du module de mémoire en fonction des images qui leur sont présentées. Ensuite, vient la phase de test durant laquelle une paire d'images est présentée au réseau, une image déjà traitée et une nouvelle image. De cette paire, le réseau détermine quelle image est déjà vue. Cette décision se base sur la comparaison du niveau d'activité des deux images. Selon la règle d'apprentissage, le modèle détermine l'image déjà vue en comparant les deux scores obtenus. En apprentissage Hebbien, l'image au score le plus élevé est considérée comme déjà vue. En apprentissage anti-Hebbien, c'est l'image au score le plus faible qui l'est.

Les premiers résultats de ce modèle furent une réplication de l'expérience de Standing (Standing, 1973). Cette expérience teste la capacité à distinguer une image vue lors de la phase d'encodage d'une image non vue. La spécificité de cette épreuve est qu'elle reprend un très grand échantillon d'images pour tester la capacité de mémoire des modèles. Plusieurs tailles d'échantillon ont été testées, chaque essai étant composé d'une phase d'apprentissage suivi du test de reconnaissance à choix-forcé. Le modèle Hebbien a montré des performances déclinantes lorsque l'échantillon dépasse 100 images, c'est-à-dire une plus grande probabilité d'erreurs lors du test de reconnaissance à choix-forcé. Le modèle anti-Hebbien, par contre,

montre des performances assez semblables aux performances humaines pour des tailles d'échantillons allant jusque 1000 (voir figure 6).

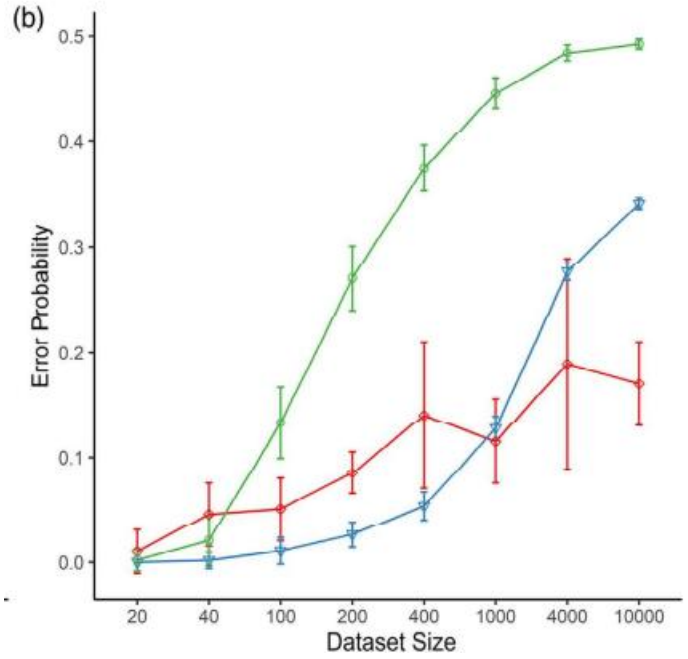

Figure 6. Probabilité d'erreur selon la taille du dataset du modèle Hebbien en vert, le modèle anti-Hebbien en bleu et l'expérience de Standing en rouge. Reproduit de Read, J., Delhaye, E., & Sougné, J. (2024). Computational models can distinguish the contribution from different mechanisms to familiarity recognition. *Hippocampus*, *34*(1), 36-50.

L'homogénéité de l'échantillon d'images a aussi été étudiée dans ces simulations. Les auteurs ont considéré 3 conditions : hétérogénéité, homogénéité faible, homogénéité forte. La condition d'hétérogénéité comprenait un échantillon d'images aléatoires repris d'une base de données. La condition d'homogénéité faible reprenait des images de chiens et la condition forte des images de chats. Ce choix se justifie par le fait que les chats présentent moins de caractéristiques distinctives que les chiens (French et al., 2001). Simplement, les chats se ressemblent plus entre eux que les chiens car ils présentent des atouts peu variables (oreilles en pointe, forme des yeux, forme de la tête).

Les résultats des simulations ont montré des performances supérieures du modèle anti-Hebbien pour toutes les conditions. Le modèle Hebbien se rapproche du hasard, c'est-à-dire un taux de décision correcte de 50%, dès que l'échantillon est très homogène même s'il est de 40 images. Le palier de 40 images était celui à partir duquel les performances du modèle Hebbien se dégradent par rapport aux performances humaines.

En résumé, le modèle anti-Hebbien semble montrer des performances comparables aux humains et conserver ses performances même en condition de forte homogénéité de l'échantillon d'apprentissage. Le modèle Hebbien, lui, semble incapable de maintenir de bonnes performances dès que l'échantillon devient homogène ou de grande taille.

Dans leurs simulations, Read et al. (2024) ont observé un effet de récence marqué dans le modèle anti-Hebbien, se traduisant par de meilleures performances de reconnaissance pour les stimuli présentés en fin d'apprentissage par rapport à ceux encodés en début de session. Ce biais temporel est absent du modèle Hebbien, dont les performances restent relativement stables quel que soit l'ordre de présentation. Ce résultat suggère que le modèle anti-Hebbien est particulièrement sensible à la nouveauté et encode un signal de familiarité de nature transitoire, fortement influencé par la récence d'exposition du stimulus. Les auteurs interprètent ce comportement comme une modélisation computationnelle de la familiarité relative. Dans cette perspective, l'apprentissage anti-Hebbien, basé sur une dépression rapide des connexions entre neurones co-activés, génère un signal de familiarité fort immédiatement après l'encodage, puis en rapide diminution lors d'expositions ultérieures à d'autres stimuli. Ce fonctionnement correspond bien à celui des neurones de récence observés dans le cortex périrhinal (Xiang & Brown, 1998), qui répondent fortement à la première présentation d'un stimulus mais dont l'activité décroît rapidement. Le modèle anti-Hebbien simulerait donc un mécanisme de détection de nouveauté fondé sur la temporalité récente, qui soutient le jugement par familiarité relative.

Les auteurs proposent que puisque l'apprentissage Hebbien fonctionne en renforçant les connexions entre les neurones co-actifs, cela signifie que à chaque fois qu'un stimulus est présenté, les connexions correspondant aux caractéristiques co-présentes dans ce stimulus sont renforcées. Cette propriété associative est étroitement liée au concept de *global matching*, un mécanisme central dans plusieurs théories de la reconnaissance, comme les *Global Matching Models* (GMM). Selon ces théories, la reconnaissance d'un stimulus dépend de la comparaison entre les caractéristiques du stimulus actuel et celles stockées en mémoire (Clark & Gronlund, 1996). Chaque stimulus en mémoire est représenté par un ensemble de caractéristiques qui sont consolidées à chaque présentation, augmentant leur force mnésique. Ce mode d'apprentissage cumulatif et distribué, s'apparente ainsi à un processus de familiarité absolue, dans lequel la force mnésique d'un stimulus est proportionnelle au nombre total d'expositions, indépendamment du contexte ou de la récence. Cette hypothèse est appuyée par des données en

IRMf montrant que le signal BOLD dans le PrC augmente avec la familiarité absolue, mesurée par la fréquence subjective des mots ((Duke et al., 2017), un profil qui reflète directement la dynamique du modèle Hebbien, dans lequel le score de familiarité augmente à mesure que les connexions se renforcent. A l'inverse, la familiarité relative se manifeste par une diminution du signal BOLD pour les stimuli récemment rencontrés, ce qui correspond au comportement du modèle anti-Hebbien, dont les scores chutent après la première exposition. Ce profil neuronal s'oppose à celui des neurones de récence évoqué plus haut et suggère l'existence de deux sous-populations neuronales au sein du PrC, l'une sensible à l'accumulation mnésique (familiarité absolue), l'autre à la nouveauté récente (familiarité relative). Dans cette perspective, les modèles Hebbien et anti-Hebbien refléteraient des mécanismes complémentaires, chacun capturant un aspect distinct de familiarité.

Suite à ces conclusions, Read et al. (2024) suggèrent que le modèle Hebbien modéliserait exclusivement la familiarité absolue et le modèle anti-Hebbien la familiarité relative. Ainsi, des performances différentielles, dépendantes du niveau de familiarité absolue initial, serait donc attendue entre les modèles. Par exemple, une différence de valeur du seuil de réponse selon les conditions, comme expliqué plus haut dans le WFME. Pour rappel, cet effet serait notamment expliqué par une différence dans le seuil de réponse selon les conditions, avec un seuil plus conservateur (plus élevé) en condition de haute familiarité absolue et signifie un besoin plus élevé de signal de familiarité pour déterminer un stimulus comme « déjà vu ». Ce seuil de réponse est illustré par la ligne pointillée sur la figure 7 proposée par Read et al. (2024). Le graphique (a), représentant une condition avec un petit jeu de données et une faible homogénéité des stimuli, propose que le modèle anti-Hebbien affiche rapidement un signal de familiarité supérieur au seuil de réponse après une exposition unique, puis diminue rapidement, illustrant une sensibilité efficace à la nouveauté contextuelle. À l'inverse, le modèle Hebbien présente une augmentation modérée et lente, restant proche du seuil critique, reflétant sa sensibilité à l'accumulation progressive d'expositions répétées. Cette augmentation modérée serait possiblement insuffisante pour une discrimination correcte dans certaines conditions, notamment en cas de haute familiarité absolue. Ce cas de figure est illustré par le graphique (b) de la figure 7. L'augmentation du signal de familiarité produite par le modèle Hebbien n'est pas suffisant pour dépasser le seuil de réponse.

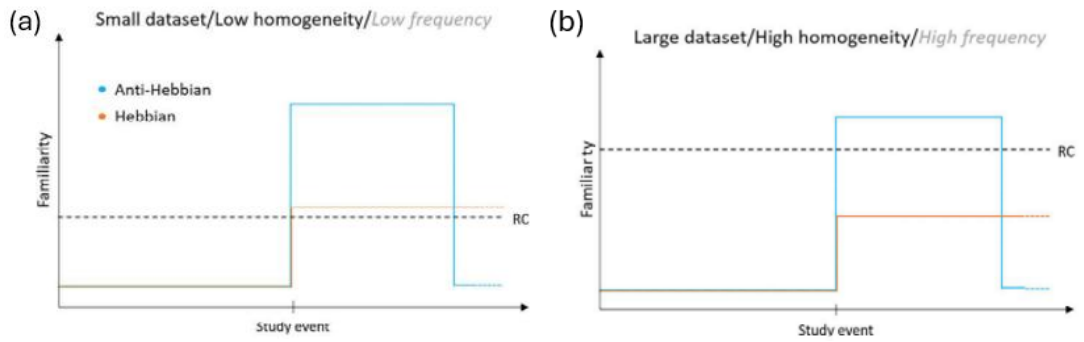

Figure 7. Représentation dynamique de la familiarité produite par le modèle Hebbien (orange) et par le modèle anti-Hebbien (bleu) en fonction des conditions expérimentales de fréquence (basse à gauche, haute à droite). Le seuil de réponse est indiqué par la ligne pointillée. Reproduit de Read, J., Delhaye, E., & Sougné, J. (2024). Computational models can distinguish the contribution from different mechanisms to familiarity recognition. *Hippocampus*, *34*(1), 36-50.

De plus, le signal de familiarité résulte probablement d'une combinaison des deux processus de familiarité (absolue et relative). L'articulation de ces processus fera l'objet d'exploration dans ce travail, la contribution de chaque processus au signal de familiarité restant à élucider.

2 Hypothèses et objectifs

La question de recherche principale occupant ce travail est la suivante : si les deux modèles représentent deux types distincts de familiarité, quelles sont les influences de différentes conditions de familiarité absolue des stimuli sur les comportements des modèles ? Une partie plus exploratoire vise à envisager des combinaisons possibles des deux modèles dans le signal de familiarité.

Sur base des analyses et hypothèses émises par Read et al. (2024), l'objectif principal de ce mémoire est d'examiner les performances des modèles Hebbien et anti-Hebbien dans différentes conditions de familiarité absolue. La distinction théorique entre la familiarité absolue (qui serait signalée dans le modèle Hebbien) et la familiarité relative (signalée dans le modèle anti-Hebbien) guide nos hypothèses spécifiques.

Une première hypothèse est que le modèle Hebbien présentera un seuil de réponse plus élevé en condition de haute familiarité comparativement à la condition de basse familiarité. Cette hypothèse découle directement de l'interprétation des graphiques de la figure 7 ainsi que du WFME. Un seuil plus élevé peut être interprété comme un besoin de plus de preuves de familiarité pour déterminer si le stimulus est « déjà vu » causé par la familiarité absolue initialement plus élevée.

Une seconde hypothèse est que le modèle Hebbien obtiendra de moins bonnes performances en condition Haute Familiarité (HF) que en condition de faible familiarité (FF). Cette prédiction est directement inspirée du WFME, selon lequel les mots à haute fréquence présentent des taux de *hits* plus faibles et des taux de *FA* plus élevés comparativement aux mots à faible fréquence.

Inversement, nous faisons l'hypothèse que le modèle anti-Hebbien restera stable et performant indépendamment des niveaux de familiarité. Ce modèle, supposé modéliser la familiarité relative grâce à son mécanisme de suppression sélective des connexions synaptiques et sa sensibilité aux différences contextuelles fines, ne devrait pas être affecté par les variations de familiarité absolue et devrait afficher de meilleures performances (indices d') que le modèle Hebbien comme suggéré par plusieurs auteurs (Bogacz & Brown, 2003a; Read et al., 2024; Tyulmankov et al., 2022).

De façon exploratoire, nous avons envisagé différentes manières de combiner les modèles Hebbien et anti-Hebbien (jusqu'ici implémenté séparément). Plus précisément, nous avons testé plusieurs pondérations des signaux de familiarité (addition, multiplication, pondération adaptative).

3 Méthodologie

Les simulations s'appuient sur l'architecture proposée par Read et al. (2024)[1], implémentée en Python version 3.10.15 (Python Software Foundation, 2023) à l'aide des bibliothèques PyTorch (Paszke et al., 2019) pour l'apprentissage profond (i.e., *deep learning*) et torchvision (Torchvision Contributors, 2023) pour la gestion des images. Deux versions distinctes du modèle ont été mises en œuvre : l'une reposant sur un apprentissage Hebbien, qui renforce les connexions entre neurones co-actifs ; l'autre sur un apprentissage anti-Hebbien, qui les affaiblit. Dans la première simulation, chaque version du modèle est testée indépendamment sur les mêmes jeux de données afin d'évaluer spécifiquement l'impact de la règle d'apprentissage sur les processus de reconnaissance par familiarité selon différents niveaux de familiarité absolue des stimuli. Dans la seconde simulation, une approche combinée des scores est appliquée *post hoc*, après exécution parallèle des modèles Hebbien et anti-Hebbien sur les mêmes images, afin d'évaluer une éventuelle complémentarité entre les deux règles.

Les scores de familiarité sont calculés et analysés à l'aide de scripts[2] en Python, utilisant notamment les bibliothèques Pandas (McKinney, 2010), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020) pour l'estimation des densités par noyau, KDE.

3.1 Architecture du modèle

Le réseau utilisé est similaire à celui exposé par Read et al. (2024) (voir figure 5). Le premier module, constitué d'un réseau profond convolutif permet l'extraction des caractéristiques d'une image. Le réseau profond utilisé dans le cadre de leur modélisation est nommé ResNet50, pré-entrainé sur ImageNet (He et al., 2016). Ce réseau est composé de 48 couches de convolutions suivies de 2 couches de *pooling*. La convolution en *deep learning* est une opération inspirée des processus visuels chez le chat (LeCun et al., 2015), qui applique un filtre à une entrée (comme une image) pour extraire des caractéristiques importantes, tandis que le *pooling* réduit la taille de l'entrée en résumant les informations, généralement en prenant la valeur maximale ou la moyenne dans chaque région du filtre. Ce réseau profond prend en entrée

[1] Ces scripts sont téléchargeables à cette adresse : https://github.com/JRead98/master
[2] Ces scripts sont téléchargeables à cette adresse : https://github.com/WilliamWar99/Master-thesis

3 matrices de valeurs qui représentent les valeurs RGB de chaque image et donne en avant-dernière couche un vecteur de 2048 valeurs, capturant les caractéristiques des images. Ce vecteur est extrait et normalisé pour avoir une moyenne de 0 et un écart-type de 1 puis deviendra la couche d'entrée du réseau *feedforward* à deux couches correspondant à la seconde étape, le module de mémoire.

Ce module de mémoire est composé de deux couches totalement connectées et prend un vecteur de 2048 valeurs en entrée et donne un vecteur de 2048 valeurs en couche de sortie. Les poids entre ces couches sont initialisés aléatoirement entre -1 et 1. C'est ce module de mémoire qui permettra un jugement de familiarité basé sur l'activité des neurones de la couche de sortie du réseau. La règle d'activation de la couche de sortie est une règle de type *m/2-winners* où la moitié des neurones dont l'activité est inférieure à la médiane de leurs scores sont considérés comme inactifs et l'autre moitié des neurones, dont l'activité est supérieure à la médiane, sont considérés actifs. L'équation (1), utilisée pour le calcul des valeurs du vecteur de sortie, est la suivante :

$$1)\ \boldsymbol{y} = \boldsymbol{W}^{\top} * \boldsymbol{x}$$

où x représente le vecteur d'entrée, $\boldsymbol{W}^{\top}$ est la transposée de la matrice des poids et x est le vecteur de sortie.

Pour représenter l'apprentissage, une modification des poids entre les couches est appliquée à chaque présentation d'un stimulus. Les neurones actifs sont ceux pour lesquels les poids vont être ajustés. Deux règles d'activations ont été implémentées. La règle d'apprentissage (Hebbienne ou anti-Hebbienne) détermine la mise à jour des poids. La règle Hebbienne se fonde sur les activations en entrée : les composantes de x dépassant la médiane sont définies comme actives. L'apprentissage Hebbien, qui augmente le poids des connexions pour les neurones actifs, se traduit par une augmentation de l'activité des neurones lors de la présentation ultérieure du même stimulus (Bogacz et al., 2001).

La mise à jour de la matrice de poids W est calculée selon l'équation (2) :

$$2)\ \Delta \mathrm{W} = \eta * \left(y_i * x_j\right)$$

où x_i équivaut au score de sortie, y_i est défini à partir de x, et vaut 1 si $x_i >$ médiane(x), sinon 0. Cela revient à renforcer les connexions des neurones fortement activés. x_j représente

les activations de la couche de ResNet50. η est le taux d'apprentissage (0.01 par défaut) qui détermine la vitesse de modification des poids.

L'apprentissage anti-Hebbien, diminue l'activité des neurones actifs en diminuant les poids des connexions entre ces neurones et les neurones de la couche d'entrée. Cela se traduit par une diminution de l'activité des neurones lors de la présentation ultérieure du stimulus appris (Bogacz & Brown, 2003a). La règle anti-Hebbienne met à jour les poids selon l'équation (3) :

$$3)\ \Delta \mathrm{W} = -\eta * \left(y_i * x_j\right)$$

où y_i est défini à partir de x, et vaut 1 si x_i > médiane(x), sinon 0. Cela revient à diminuer les connexions des neurones fortement activés. x_j représente les activations de la couche de ResNet50. η est le taux d'apprentissage (0.01 par défaut) qui détermine la vitesse de modification des poids. Le signe négatif de $-\eta$ permet de diminuer les connexions des neurones fortement activés.

La mise à jour est ensuite appliquée à la matrice de poids par addition directe selon l'équation (4) :

$$4)\ w_{ij(t+1)} = w_{ij(t)} + \Delta w_{ij}$$

Les mises à jour des poids sont effectuées une seule fois par image durant la phase d'apprentissage, sans répétition, conformément à la méthodologie définie par Read et al. (2024).

3.2 Validation des conditions de familiarité

Les différentes conditions de familiarité reposent sur une sélection et une manipulation d'images, issues d'un grand jeu de données existant, pour obtenir trois jeux de données distincts selon leur niveau de familiarité absolue.

3.2.1 Origine des images

Les stimuli visuels utilisés dans les simulations proviennent du set de données ImageNet-mini[3] (Ifigotin, 2021) un sous-ensemble léger du corpus ImageNet original (Deng et al., 2009). ImageNet Mini contient 1000 classes d'objets, réparties en deux sous-ensembles : un *training set* et un *validation set*. Le réseau ResNet50 utilisé comme encodeur visuel dans le modèle a été pré-entraîné sur le *training set* complet d'ImageNet ; le *validation set* est utilisé dans le but de contrôler la capacité de ResNet50 à catégoriser correctement de nouvelles images non vues. Lors de cette validation, le réseau convolutif est fixé et ne se modifie plus, il n'apprend donc pas ces images. En d'autres termes, il s'agit d'images qui n'ont pas été apprises par le modèle – via une modification de ses poids – mais dont la catégorie a bien été apprise lors de son entraînement.

3.2.2 Conditions expérimentales

Trois niveaux de familiarité implicite ont été définis selon 1) l'origine des images dans le dataset et 2) une transformation effectuée *à posteriori*. Un exemple de chaque jeu de données est représenté sur la figure 8.

Premièrement, le jeu de données Haute Familiarité (HF) est constitué d'images issues du *training set* d'ImageNet, et donc supposées avoir été déjà vues par le modèle ResNet50 durant sa phase de pré-entraînement. Nous partons donc du principe que ces images présentent un haut niveau de familiarité *absolue*.

Le deuxième jeu, nommé Faible Familiarité (FF) contient des images issues du *validation set* d'ImageNet, ce qui garantit qu'elles n'ont jamais été apprises par ResNet50. Nous partons du principe qu'elles présentent un niveau de familiarité *absolue* faible, mais conservent une structure perceptive et sémantique (i.e., la catégorie de l'image) intacte.

Troisièmement, le dernier jeu de données, nommé Non Familiarité (NF), contient les mêmes images que dans la condition de FF, mais dont la structure visuelle a été altérée pour les rendre méconnaissables par ResNet50. Autrement dit, nous partons du principe que ce dataset ne présente aucune forme de familiarité pour le modèle. Ces images subissent une

[3] Les jeux de données sont téléchargeable à cette adresse :https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000

transformation[4] de Fourier par blocs : chaque image est divisée en 25 blocs (5×5) et une transformée de Fourier est appliquée à chaque bloc. La phase spectrale est aléatoirement modifiée, puis l'image est reconstruite. Cette manipulation préserve l'amplitude fréquentielle (et donc la luminance globale), mais détruit les régularités spatiales fines, rendant le traitement perceptif difficile pour le modèle (Geirhos et al., 2022).

Figure 8. Images d'une même catégorie (mastiff tibétain) pour chaque dataset (gauche à droite : HF, FF, NF)

3.2.3 Méthode de sélection

Pour maximiser le contraste entre les conditions, une sélection rigoureuse a été opérée à partir des scores donnés par ResNet50 (pré-entraîné). Toutes les images du *training* et du *validation set* ont été passées à travers ResNet50. En sortie, ResNet50 donne un vecteur softmax qui contient des valeurs entre 0 et 1 et dont la somme vaut 1. Chaque score softmax donné par ResNet50 représente une prédiction que l'image appartienne à une certaine classe. Ce score peut être interprété comme la confiance du réseau que l'image appartienne à la classe correspondante, avec le choix de reconnaissance associé au score softmax maximum.

Pour chaque classe, l'image du *training set* ayant le score softmax maximum le plus élevé a été sélectionnée pour le set HF. L'image du *validation set* ayant le score le plus faible a été sélectionnée pour le set FF. Cette même image FF a été transformée pour constituer le set NF. Aucune de ces images NF transformées n'a été correctement reconnue par ResNet50 par la suite. Ce protocole permet de garantir que le modèle encode différemment chaque condition selon son degré d'exposition antérieure aux images et la structure perceptive disponible. Cette sélection renvoie un total de 997 images pour le set HF et 965 images pour FF et NF. Pour chaque run de chaque simulation, un échantillon aléatoire de 100 images est tiré dans le set concerné pour constituer l'ensemble d'apprentissage.

[4] Le script est disponible à cette adresse : https://martin-hebart.de/webpages/code/stimuli.html

3.3 Simulation 1 : Effet de la familiarité absolue

3.3.1 Objectif

Cette première simulation vise à examiner comment la familiarité absolue, induite par la nature perceptive et statistique des images, influence les mécanismes de reconnaissance fondés sur la familiarité. Plus précisément, l'objectif est d'évaluer si le seuil de reconnaissance varie en fonction du niveau de familiarité absolue et si cette modulation dépend de la règle d'apprentissage utilisée (Hebbienne ou anti-Hebbienne) ainsi que d'évaluer si la sensibilité perceptive du modèle est également affectée par le niveau de familiarité absolue des images.

3.3.2 Protocole expérimental

Chaque exécution s'articule en deux étapes : une phase d'apprentissage suivie d'une phase de détermination des scores qui s'apparente à une phase de test.

Lors de la phase d'apprentissage, le modèle est exposé à un ensemble de 100 images sélectionnées aléatoirement parmi le *dataset* cible (HF, FF, NF). Ces images sont normalisées (redimensionnement à 224x224 pixels, standardisation ImageNet) puis transmises au modèle une par une. Le processus repose sur un apprentissage non supervisé, où la mise à jour des poids s'effectue sans signal de correction. Chaque image est traitée en deux étapes. D'abord l'extraction des représentations visuelles dans laquelle l'image est encodée par ResNet50, utilisé ici comme un module d'extraction de caractéristiques. Le vecteur de sortie de l'avant dernière couche (dimension 2048) est normalisé par soustraction de la moyenne et division par l'écart-type de ses composantes (z-scoring). Ensuite vient l'étape de mise à jour des poids du module de mémoire durant laquelle le vecteur extrait est projeté dans le module de mémoire.

L'entrainement se déroule en un seul passage sur chacune des 100 images. En fin d'apprentissage, les poids finaux sont conservés pour la phase de détermination des scores.

Lors de la phase de détermination des scores et pour chaque exécution, le modèle traite 200 images : 100 images vues lors de la phase d'apprentissage (désignées ici comme images X ou « old ») et 100 nouvelles images (désignées ici comme images Z ou « new »). Pour chaque image, un score de familiarité (d(x) ou d(z)) est calculé à partir de l'activité des neurones de la couche de sortie du modèle.

Selon la règle d'apprentissage utilisée, le score de familiarité est défini comme suit : pour l'apprentissage Hebbien, le score est obtenu par un produit scalaire entre le vecteur

d'entrée x_i (représentation issue de ResNet50) et l'activité de sortie y_i du module de mémoire selon l'équation (5) :

$$5)\ d(x) = x_i * y_i$$

Pour l'apprentissage anti-Hebbien, l'activité des neurones est d'abord dichotomisée selon leur médiane suivant l'équation (6) :

$$6)\ y_j = 1\ si\ y_j > médiane(y), sinon\ y_j = 0$$

Puis le score est obtenu par l'équation (7) :

$$7)\ d(x) = \sum_j a_j * y_j - \sum_j \left(1 - a_j\right) * y_j$$

Ce score correspond à la différence d'activation entre les neurones les plus actifs et les moins actifs.

Les 200 scores sont alors répartis en deux distributions. d(x) contient les scores des images « old » et d(z) les scores des images « new ».

3.3.3 Analyse des performances

L'analyse des performances vise à évaluer la capacité du modèle à discriminer les images déjà vues (« old ») de celles qui sont nouvelles (« new »), sur la base des scores de familiarité extraits après la phase d'apprentissage. Cette évaluation repose sur un paradigme de reconnaissance oui/non dans lequel la réponse du modèle est inférée sur base du seuil optimal entre les distributions. Les analyses peuvent être réalisées après deux étapes. D'abord, la phase de détermination du seuil optimal prend place, pour chaque run, les deux distributions d(x) et d(z) sont lissées à l'aide d'une estimation par noyau gaussien (Kernel Density Estimation, KDE) (Parzen, 1962). L'utilisation d'une estimation par noyau permet d'obtenir une approximation continue et lissée de la densité de probabilité des scores (voir figure 9), ce qui est crucial lorsque les distributions présentent des recouvrements partiels.

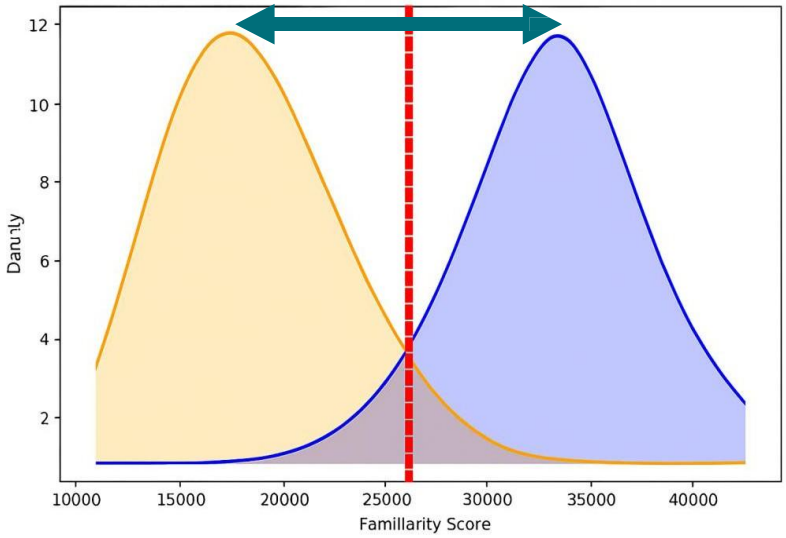

Figure 9. Distribution lissée par estimation par noyau (KDE) des scores de familiarité en apprentissage anti-Hebbien. En jaune, d(x), les scores des images « old », en bleu, d(z) les scores des images « new ». Le seuil est représenté par la ligne verticale pointillée rouge et la séparation des distributions par la double flèche horizontale turquoise.

Le seuil optimal de familiarité θ est défini comme le point d'intersection entre les courbes KDE le plus proche de la moyenne des médianes des deux distributions par l'équation (8) choisie de façon arbitraire pour déterminer le seuil sur base d'un parallèle avec la théorie de détection du signal qui place le Criterion neutre à l'intersection des courbes (cf. figure 1) :

$$8)\ \theta = \arg \min_{t \text{ tel que } \mathrm{KDE}_{d(x)}(t)=\mathrm{KDE}_{d(z)}(t)} \left| t - \frac{\text{mediane}(d(x))+\text{mediane}(d(z))}{2} \right|$$

où arg min est un opérateur mathématique qui renvoie la valeur de t minimisant l'expression entre accolades. t est une variable sur l'axe d'abscisse des scores de familiarité et représente un candidat potentiel pour le seuil. L'expression en indice de arg min exprime la condition d'égalité des courbes de densité de d(x) et d(z). L'expression entre accolade assure que t se situe entre les médianes des deux distributions.

Ce seuil est ensuite utilisé pour inférer un comportement de type « oui/non » : une image est classée comme « old » si son score de familiarité dépasse (ou est inférieur à, en anti-Hebbien) le seuil lors de la phase de calcul de l'indice de séparation des distributions. Afin d'estimer la capacité du modèle à différencier les images « old » des images « new », un indice de séparation entre les deux distributions de scores est calculé. Cet indice ne correspond pas

exactement au d' de la théorie de détection du signal, mais repose sur une mesure de distance normalisée entre moyennes (9) :

$$9)\ \Delta = \frac{\mu_{d(x)} - \mu_{d(z)}}{\sqrt{\frac{\sigma^2_{d(x)} + \sigma^2_{d(z)}}{2}}}$$

où $\mu_{d(x)}$ et $\mu_{d(z)}$ désignent respectivement les moyennes des scores de familiarité des images apprises et nouvelles, et σ^2 leur variance. Cette mesure continue reflète la séparabilité entre les deux distributions et est utilisée comme indicateur de performance dans les comparaisons entre règles d'apprentissage et niveaux de familiarité. La formule (12) décrite plus bas du d' canonique de la théorie du signal (Green & Swets, 1966) n'est pas applicable dans ce contexte car les réponses binaires (« old »/ « new ») pour chaque image n'ont pas été extraites lors des simulations.

3.4 Simulation 2 : Méthode d'intégration des modèles

Cette seconde simulation vise à tester l'hypothèse selon laquelle les familiarités absolue et relative, modélisées respectivement par les règles d'apprentissage Hebbienne et anti-Hebbienne, pourraient être intégrées pour générer un signal de reconnaissance unique semblable au performances humaines. L'objectif est donc d'évaluer si une combinaison des deux modèles rapproche les performances des comportements humains.

3.4.1 Protocole expérimental

Le protocole expérimental de la simulation 2 est identique à celui de la simulation 1, à ceci près que les deux modèles (Hebbien et anti-Hebbien) sont exécutés en parallèle sur les mêmes données. Pour chaque run, un ensemble de 100 images d'apprentissage est sélectionné aléatoirement dans le dataset cible (HF, LF ou NF), et les mêmes images dans le même ordre sont présentées aux deux modèles. Cette contrainte permet de comparer et combiner les scores de familiarité de manière strictement équitable.

La phase de test repose sur 200 images (100 anciennes et 100 nouvelles), identiques pour les deux modèles. Pour chaque image X (« old ») ou Z (« new »), on extrait un score

$S_H(X)\ ou\ S_H(Z)$ issu du modèle Hebbien et un score $S_{AH}(X)\ ou\ S_{AH}(Z)$ issu du modèle anti-Hebbien. Ces scores sont ensuite combinés pour produire un score unifié de familiarité selon différentes méthodes décrites ci-dessous.

3.4.2 Méthodes d'intégration des modèles

Plusieurs méthodes de combinaisons sont testées. D'abord, les combinaisons simples selon les opérations d'addition par l'équation (10) et de multiplication par l'équation (11) :

$$10)\quad S_{\text{add}}(X) = S_H(X) + S_{AH}(X)$$

$$11)\quad S_{\text{mult}}(X) = S_H(X) * S_{AH}(X)$$

où S_H est le score donné par le modèle Hebbien et S_{AH} est le score donné par le modèle anti-Hebbien.

Ces règles supposent une contribution égale de chaque modèle, sans ajustement dynamique en fonction de leur performance relative.

Une méthode supplémentaire de combinaison adaptative est choisie pour sa pondération selon la variance des distributions. Cette méthode repose sur le principe que chaque modèle (Hebbien ou anti-Hebbien) contribue au score final proportionnellement à la fiabilité de son signal : plus une distribution des scores est précise (c'est-à-dire avec une faible variance), plus son estimation de familiarité est considérée comme fiable et influente dans le score combiné.

Ce type de pondération s'inspire des principes exposés dans les modèles à systèmes complémentaires proposés par Norman & O'Reilly (2003) et développés plus avant par O'Reilly et al. (2014). Dans ces travaux, bien qu'aucune formule explicite de combinaison pondérée par la variance ne soit donnée, les auteurs soulignent que le poids relatif des contributions d'un système dépend de la robustesse et de la fiabilité de son signal. Ils discutent par exemple de la variance du signal de familiarité pour estimer une distinction entre cibles et leurres, et adaptent dynamiquement la contribution de chaque module selon le contexte ou les propriétés du stimulus. Ainsi, l'idée de fonder la décision finale sur la précision relative des sources mnésiques est bien présente dans leur cadre théorique, même si leur approche repose plutôt sur des mécanismes *winner-take-all* ou sur l'ajustement du seuil.

La méthode adoptée ici formalise cette intuition de manière plus explicite, en utilisant une règle pondérée. Le score de familiarité combiné est calculé par cette équation (12) :

$$12) \quad S_{\text{comb}}(i) = \frac{p_H * S_H(i) + p_{AH} * S_{AH}(i)}{p_H + p_{AH}}, \quad \text{où } p_k = \frac{1}{\sigma_k^2}$$

Où p_k sont les variances des scores pour les images « old » ou « new » (sur l'ensemble du run) pour chacun des modèles. p_H et p_{AH} sont les pondérations attribuées aux scores Hebbien et anti-Hebbien, respectivement, selon leur précision. Cette stratégie vise à modéliser une intégration optimale de deux signaux complémentaires, en tenant compte de leur bruit respectif.

3.4.3 Analyses des performances

L'analyse des performances de la Simulation 2 vise un objectif similaire à celui de la Simulation 1 et repose sur le même paradigme de reconnaissance oui/non avec d'abord une phase de détermination du seuil. Pour chacune des méthodes de combinaison (addition, multiplication, pondération), un seuil optimal est calculé à partir des distributions des scores combinés pour les images « old » et « new » (d(x) et d(z)). La procédure est identique à celle décrite dans la simulation 1 : les deux distributions sont lissées par estimation de densité par noyau (KDE), et le seuil θ est défini comme le point d'intersection entre les deux courbes de densité le plus proche de la moyenne de leurs médianes selon l'équation (7). Une fois ce seuil défini, la « réponse » du modèle est déterminée comme suit : si $S_{\text{comb}}(i) > \theta$, l'image est classée comme « old », sinon, elle est classée comme « new ».

Vient ensuite la phase de calcul de l'indice d'. Contrairement à la Simulation 1, qui s'appuyait directement sur les valeurs des seuils, la Simulation 2 permet d'obtenir des réponses catégorielles (« old »/ « new ») pour chaque image à partir du score combiné et du seuil. Cela permet une évaluation selon les principes classiques de la Signal Detection Theory (SDT), via les quatre catégories : *Hit*, *Miss*, *FA, CR*.

A partir de ces données, on calcule pour chaque méthode de combinaison deux taux. Les taux de *Hits* et de *FA*, identifiant respectivement la proportion d'images « old » correctement reconnues et la proportion d'images « new » incorrectement reconnues, sont calculés par les formules (13) et (14) :

$$13) \quad pHit = Hits/(Hits + Miss)$$

$$14) \quad pFA = FA/(FA + CR)$$

Ces proportions permettent le calcul du d' classique (Green & Swets, 1966), défini comme la formule (15) :

$$15) \quad d' = Z(\text{pHit}) - Z(\text{pFA})$$

où Z désigne la transformation inverse d'une distribution normale cumulée. L'interprétation de la valeur du d' se fait selon plusieurs intervalles. De 0,5 à 1, la discrimination est faible à modérée. De 1,5 à 2, elle est modérée à forte et au-dessus de 2, la discrimination est dite excellente (Green & Swets, 1966).

4 Résultats

Cette section présente les résultats des simulations visant à évaluer les effets de la familiarité absolue sur les performances des modèles Hebbien et anti-Hebbien, ainsi que les effets des différentes combinaisons de scores.

4.1 Simulation 1 : effet de la familiarité absolue

4.1.1 Objectif

Cette simulation avait pour objectif d'évaluer comment différents niveaux de familiarité absolue influencent les comportements et performances de reconnaissance des modèles dotés d'un apprentissage Hebbien ou anti-Hebbien. Trois conditions de familiarité ont été testées : la condition HF qui contient des images hautement familières pour le modèle (issues du *training set* de ResNet50), la condition FF contenant des images jamais vues mais perceptivement intactes (*validation set* de ResNet50) et enfin, la condition NF avec les mêmes images que FF, mais avec dégradation perceptive.

Les performances ont été analysées à partir du seuil de familiarité optimal calculé pour chaque run, ainsi que d'un indice de séparation des distributions de familiarité pour les images « old » (vues pendant la phase d'apprentissage) et « new » (non vues).

4.1.2 Seuil de familiarité

Pour chaque run indépendant (100 runs par condition), un seuil de reconnaissance θ a été déterminé à partir des distributions des scores de familiarité « old » et « new ». Ce seuil correspond au point d'intersection entre les densités estimées de d(x) et d(z), le plus proche de la moyenne de leurs médianes.

Les seuils, moyennés sur 100 runs, obtenus dans chaque condition sont représentés sur la figure 10.

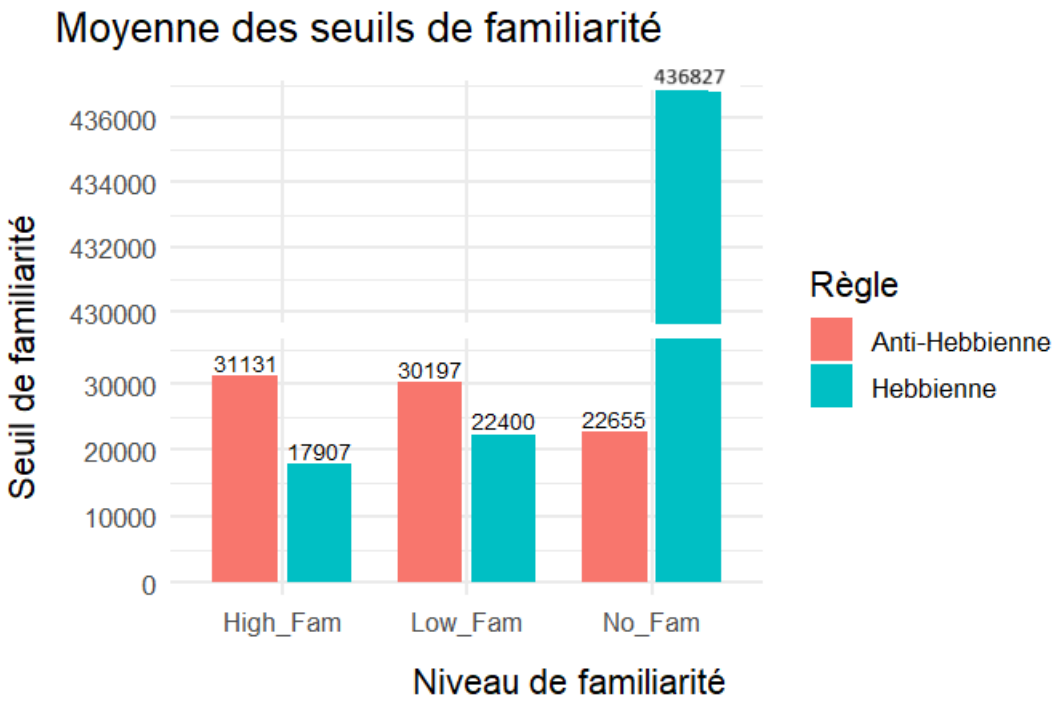

Figure 10. Seuils moyens de familiarité en fonction du type de règle d'apprentissage (Hebbienne et anti-Hebbienne) et du niveau de familiarité absolue (HF, FF, NF). L'axe des Y est tronqué entre 30 000 et 430 000.

Bien que les règles d'apprentissage Hebbienne et anti-Hebbienne entrainent des dynamiques inversées dans la manière dont les scores de familiarité évoluent, les seuils affichés ici sont déterminés selon une méthode identique pour les deux modèles. Il s'agit du point d'intersection des distributions de familiarité pour les images « old » et « new », indépendamment du sens du changement. Par conséquent, les seuils peuvent uniquement être comparés au sein de chaque modèle, mais le sens du score de seuil ne doit pas être interprété de la même manière entre les deux approches. En Hebbien, une augmentation signifie un besoin de plus de preuves pour reconnaitre une image, alors qu'en anti-Hebbien, c'est une diminution du seuil qui désigne un besoin de plus de preuves pour reconnaitre une image.

On observe des profils contrastés entre les deux types d'apprentissage. Le modèle Hebbien montre une augmentation du seuil lorsque la familiarité absolue diminue (HF → FF → NF), avec une augmentation très marquée en NF. Pour le modèle anti-Hebbien, le seuil reste stable entre les conditions HF et FF, mais diminue légèrement en NF.

La figure 11 présente les distributions des seuils de familiarité (θ) obtenus à partir de 100 runs pour chaque combinaison de règle d'apprentissage (Hebbienne et anti-Hebbienne) et de niveau de familiarité (HF, FF, NF). Les courbes correspondent à des estimations de densité

par noyau (KDE) et la ligne pointillé rouge indique la moyenne des seuils pour chaque condition.

Les écarts-types des seuils de familiarité sont les suivants : en HF, σ = 526,32 pour le modèle Hebbien (graphique a) et σ = 557,36 pour le modèle anti-Hebbien (graphique d) ; en FF, σ = 613,86 pour le modèle Hebbien (graphique b) et σ = 441,01 pour le modèle anti-Hebbien ; enfin, en NF, σ = 43594,56 pour le modèle Hebbien et σ = 348,32 pour le modèle anti-Hebbien.

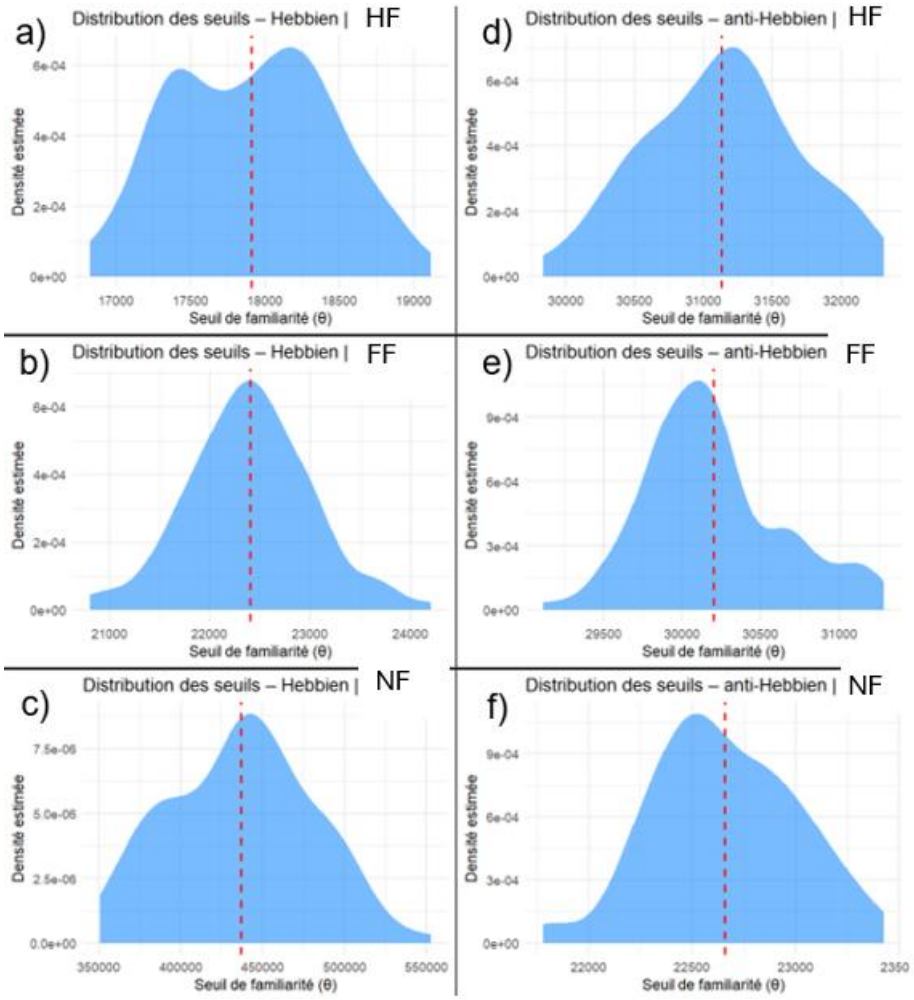

Figure 11. Graphique de distributions des seuils sur 100 runs. La ligne pointillée rouge représente le seuil moyen. A gauche : apprentissage Hebbien, à droite : anti-Hebbien. De haut en bas : HF, FF, NF.

4.1.3 Indice de séparation des distributions

Pour évaluer la capacité discriminative du modèle indépendamment du seuil, un indice de séparation a été calculé. Il correspond à la distance normalisée entre les moyennes des scores d(x) et d(z), reflétant le degré de séparation des distributions de familiarité. Les résultats moyens sont présentés ci-dessous (Tableau 1) :

Tableau 1

Indices de séparation des distributions de scores de familiarité selon la condition de familiarité et la règle d'apprentissage

	Indice de séparation HF (écart-type)	**Indice de séparation FF (écart-type)**	**Indice de séparation NF (écart-type)**
Hebbienne	3,29 (0,40)	2,72 (0,27)	0,10 (0,13)
Anti-Hebbienne	-5,82 (0,56)	-5,08 (0,39)	-1,54 (0,14)

Ces résultats révèlent plusieurs éléments importants. D'abord, le modèle Hebbien affiche une meilleure séparation en HF. Toutefois, cette performance chute lorsque les images deviennent moins familières (FF). La condition NF, basée sur les images perceptivement altérées, affiche un indice proche de 0, indiquant une quasi-confusion des deux catégories. Ensuite, le modèle anti-Hebbien montre des performances très élevées pour les conditions HF et FF. La condition NF semble altérer sa performance mais l'indice de séparation indique tout de même une discrimination efficace et au-dessus du niveau du hasard.

4.2 Simulation 2 : Méthodes d'intégration des modèles

4.2.1 Objectif

Après avoir examiné séparément les contributions des règles d'apprentissage Hebbienne et anti-Hebbienne, cette seconde simulation explore la possibilité d'intégrer ces deux signaux de familiarité au sein d'un score unifié. L'objectif est d'évaluer si une telle combinaison permet de se rapprocher des performances humaines. Plusieurs méthodes d'intégration des scores ont été testées : l'addition des scores de familiarité, la multiplication des scores de familiarité, et une pondération adaptative en fonction de la variance des distributions.

Les performances ont été mesurées à l'aide de la SDT : pour chaque image, le score combiné est comparé à un seuil, et la réponse (old/new) est classée comme *Hit*, *Miss*, *FA* ou *CR* en référence au seuil. A partir de cela, un indice d' classique a été calculé.

4.2.2 Seuil de familiarité

Pour chaque run (n=100 par condition), un seuil optimal a été déterminé à partir de l'intersection des distributions de scores (Hebbien et anti-Hebbien combinés) « old » et « new », estimées par noyau (KDE). Ce seuil représente la valeur à partir de laquelle le modèle bascule sa décision entre « déjà vu » et « nouveau ». Le tableau 2 ci-dessous rapporte les moyennes et écart-types de seuils calculés pour chaque méthode de combinaison (voir figure 12 pour les graphiques) :

Tableau 2

Moyennes et écarts-types des seuils de familiarité pour chaque méthode de combinaison selon la condition de familiarité

	Seuil moyen addition (écart-type)	**Seuil moyen multiplication (écart-type)**	**Seuil moyen pondération (écart-type)**
HF	36902,56 (475,95)	338441461,6 (9013784,38)	18872,94 (231,86)
FF	41442,9041 (661,93)	423838797 (12303812)	18923,8294 (207,12)
NF	459763,54 (45516,0018)	8032413939 (904330899)	18123,1426 (201,27)

On observe des profils très contrastés entre les méthodes de combinaison. Les méthodes additive et multiplicative produisent des seuils de plus en plus élevés à mesure que la familiarité diminue, avec une explosion du seuil en NF, particulièrement pour la multiplication qui atteint des valeurs extrêmes. En revanche, la méthode pondérée, qui ajuste la contribution de chaque modèle en fonction de la variance de ses scores, génère des seuils beaucoup plus stables et d'amplitude modérée, avec peu de variabilité inter-run.

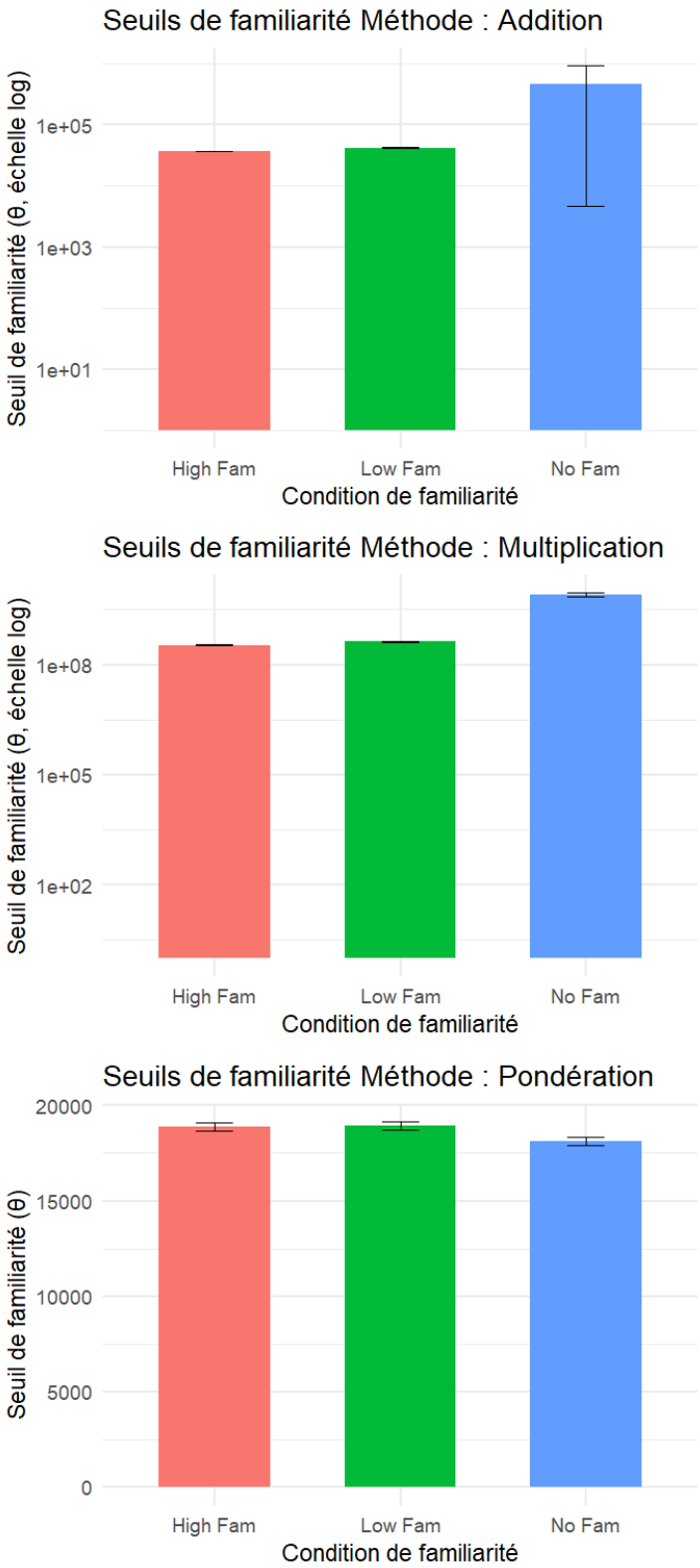

Figure 12. Seuils moyens et écarts-types selon la condition (HF en rouge, FF en vert et NF en bleu). Les échelles en ordonnées sont logarithmiques pour les méthodes additives et multiplicatives.

4.2.3 Performance de reconnaissance (indice d')

Les performances de reconnaissance ont été évaluées à l'aide de l'indice d', une mesure standard de la sensibilité du système à discriminer les images « old » des images « new » à partir des « décisions » catégorielles simulées. Pour chaque run, le taux de *Hit* et de *FA* ont été extraits, puis transformés en scores d' via la fonction Z de la loi normale cumulative. Les moyennes obtenues sont présentées ci-dessous par le tableau 3 et la figure 13.

Tableau 3

Moyennes et écarts-types des indices d' par méthode de combinaison selon la condition de familiarité

	d' moyen addition (écart-type)	d' moyen multiplication (écart-type)	d' moyen pondération (écart-type)
HF	3,79 (0,44)	3,74 (0,42)	1,05 (0,19)
FF	2,86 (0,26)	2,80 (0,34)	1,15 (0,22)
NF	0,10 (0,18)	-0,12 (0,19)	1,49 (0,22)

Tableau 4. Moyennes et écarts-types des indices d' par méthode de combinaison et niveau de familiarité implicite.

Ces résultats révèlent un phénomène intéressant. En HF et FF, les méthodes additive et multiplicative surpassent nettement la pondération adaptative, suggérant qu'une simple intégration linéaire des signaux suffit lorsque les représentations perceptives sont intactes. En NF, les indices d' chutent fortement en addition et en multiplication, ce qui indique une confusion du modèle dans cette condition perceptive dégradée. Par contre, la méthode pondérée atteint ici une meilleure performance relative, dépassant nettement les deux autres. Cela suggère que dans des conditions bruitées ou perturbées, pondérer les contributions selon leur fiabilité améliore la capacité du modèle à discriminer les images vues des nouvelles.

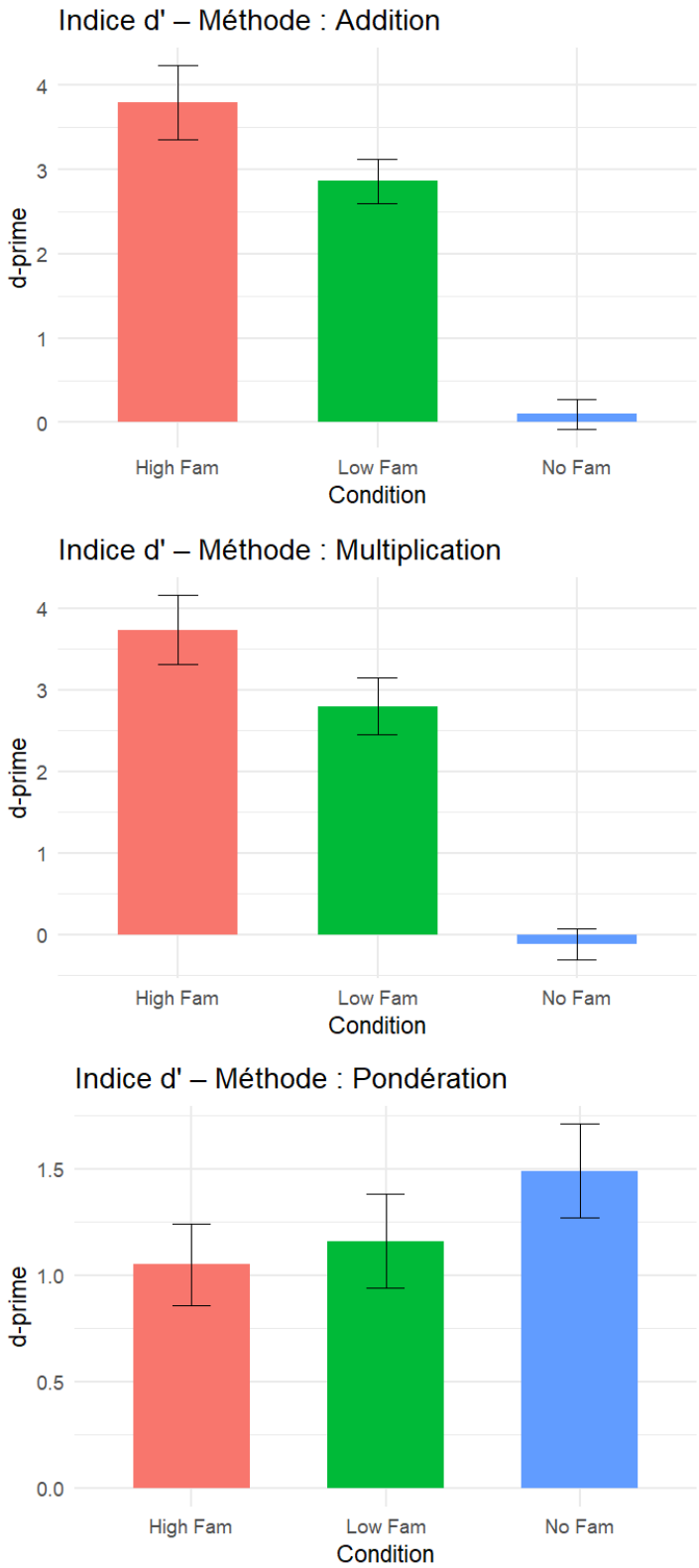

Figure 13. Indices d' moyens et écarts-types selon la condition (HF en rouge, FF en vert, NF en bleu).

5 Discussion

Ce travail avait pour objectif de modéliser deux formes distinctes de familiarité (absolue et relative) à l'aide de règles d'apprentissage simples appliquées à un réseau de neurones artificiels. En s'inspirant grandement du modèle proposé par Read et al. (2024), deux versions du même système ont été mises en œuvre. L'une utilisait une règle Hebbienne, censée capturer la familiarité absolue en renforçant les représentations précédemment activées. L'autre, anti-Hebbienne, modélisait la familiarité relative en inhibant les configurations déjà rencontrées.

L'objectif principal de la première simulation était d'évaluer l'influence du niveau de familiarité absolue des images (HF, FF, NF) sur la capacité du modèle à reconnaitre les stimuli appris, et de déterminer si cette modulation dépendait de la règle d'apprentissage utilisée. Il était attendu que la familiarité absolue, encodée par le modèle Hebbien, soit particulièrement sensible aux niveaux de familiarité élevés, tandis que la règle anti-Hebbienne offrirait une plus grande robustesse.

La seconde simulation constituait une tentative exploratoire d'intégration des deux formes de familiarité selon plusieurs méthodes. Les méthodes testées étaient l'addition, la multiplication ou une pondération adaptative fondée sur la fiabilité de chaque signal.

La discussion qui suit propose une interprétation des résultats obtenus à la lumière de la littérature, en confrontant les profils observés aux prédictions théoriques. Elle s'attarde également sur les apports, les limites du modèle et les perspectives ouvertes par cette approche computationnelle de la reconnaissance fondée sur la familiarité.

5.1 Effet de la familiarité absolue

L'hypothèse initiale prévoyait un seuil de réponse plus élevé pour le modèle Hebbien en condition de haute familiarité (HF) qu'en faible familiarité (FF), sur la base de l'idée qu'une condition de haute familiarité absolue exigerait un niveau de preuve plus important pour éviter les fausses reconnaissances (cf. WFME). Les résultats obtenus contredisent cette prédiction : le seuil en HF est plus bas qu'en FF. Ce profil suggère qu'un niveau élevé de familiarité absolue entraîne au contraire un biais décisionnel plus « libéral », le système étant plus enclin à classer

un stimulus comme « déjà vu » même avec une évidence plus faible. Cependant, l'interprétation des seuils dans ce travail s'éloigne de celle de la SDT (Macmillan & Creelman, 2004). En effet, le criterion de la SDT est un indice de biais de réponse par rapport au seuil optimal. Ici, le seuil est déterminé par une formule qui vise directement le seuil optimal et n'exprime donc pas de biais de réponse comme il est possible d'en voir dans l'expérimentation humaine. Pour cela, il est difficile de comparer les seuils avec les valeurs de criterion observées dans la littérature. Cependant, le criterion représente également la valeur du signal représentant un seuil, une frontière entre deux distributions (Green & Swets, 1966). Dans cette optique, la comparaison entre les seuils au sein des modèles pourrait être valide et pertinente. Le seuil plus faible en HF est donc inverse aux prédictions faites par le WFME (Coane et al., 2011; Reder et al., 2000). De plus, le WFME peut s'expliquer par des différences de capacité de discrimination (d') selon les conditions, avec un d' supérieur en condition de faible fréquence du mot. Nos données ne vont pas dans ce sens non plus. Le d' en HF est supérieur à celui en FF.

Cette double discordance de nos données avec le WFME pourrait s'expliquer de plusieurs façons. Premièrement, comme son nom l'indique, le WFME est un effet initialement observé dans des expérimentations sur des mots. Cependant, certains auteurs suggèrent un effet similaire sur les images (Snodgrass et al., 1974; Snodgrass & Burns, 1978; Snodgrass & McClure, 1975). Ensuite, la façon dont la fréquence est manipulée dans ce travail diffère des manipulations traditionnelles (Joordens & Hockley, 2000; Reder et al., 2000). Notre manipulation repose sur une sélection *a priori* d'images vues ou non par ResNet50 lors de son entrainement. Les images vues et non vues appartiennent aux mêmes classes et sont donc sémantiquement très similaires. Hors, plusieurs auteurs ont proposé que le WFME ne repose pas uniquement sur des différences de fréquence de mots mais également sur des facteurs sémantiques (Monaco et al., 2007; Popov & Reder, 2024). Les auteurs soulignent que les effets de fréquence ne sont pas purement perceptifs ou statistiques, mais reposent sur des différences qualitatives dans les représentations sémantiques : les concepts fréquents partagent davantage de traits avec d'autres concepts, augmentant le « bruit contextuel » lors de la récupération, alors que les concepts rares activent des représentations plus spécifiques. Dans ce cadre, notre manipulation HF et FF, pour laquelle les images appartiennent à des classes sémantiques identiques, pourrait ne pas induire le même type de séparation sémantique que dans les expériences verbales classiques. Si les images HF et FF mobilisent des réseaux de traits perceptifs et sémantiques largement superposés, ResNet50 pourrait extraire des caractéristiques similaires et entrainer un signal de familiarité

relativement similaire par le modèle Hebbien, expliquant l'absence d'effet attendu de type WFME.

Une seconde hypothèse postulait que le modèle anti-Hebbien, supposé modéliser la familiarité relative, serait relativement insensible au niveau de familiarité absolue (HF, FF, NF), maintenant un seuil de décision stable quelles que soient les conditions. Les résultats confirment partiellement cette prédiction : bien que le seuil ne varie pas de manière marquée entre HF et FF, une diminution notable est observée dans la condition NF, indiquant que la familiarité relative reste sensible à des altérations perceptives fortes, tout en appuyant le parallèle possible entre modèle anti-Hebbien et familiarité relative lorsque les stimuli sont « naturels ».

Les effets de la condition NF sur les deux modèles sont possiblement compréhensibles en abordant les travaux sur la robustesse des réseaux convolutifs (e.g. ResNet50) qui montrent que ce type d'altération fréquentielle, comme la transformée de Fourier, peut entrainer une chute drastique des performances lorsque le réseau s'appuie sur des indices de texture (Hendrycks & Dietterich, 2019; Yin et al., 2020). Dans notre cadre, la chute de performance des modèles est claire. Le modèle Hebbien montre des performances proches du hasard et le modèle anti-Hebbien des performances clairement plus faibles que dans les autres conditions, malgré qu'elles soient toujours modérées. Ceci s'expliquerait par le fait que les réseaux de neurones artificiels semblent plus sensibles au bruit blanc que les humains, par contre, quand le bruit floute la forme globale, ce qui est le cas ici, les humains sont plus affectés comparativement au réseau (Jang et al., 2021).

La sens de la différence de performances entre les modèles correspond au résultat attendu. Dans la littérature, plusieurs modèles reposant sur un apprentissage anti-Hebbien ont montré des capacités de reconnaissance supérieures aux modèles Hebbiens (Bogacz & Brown, 2003; Read et al., 2024; Tyulmankov et al., 2022).

5.2 Intégration des modèles

La simulation 2 visait à explorer la possibilité que les signaux issus des deux règles d'apprentissage puissent être intégrés dans un mécanisme unifié de reconnaissance. Cette démarche s'appuie sur l'hypothèse selon laquelle ces deux formes de familiarité, bien que

fondées sur des dynamiques computationnelles distinctes, pourraient coexister et se combiner pour produire un signal global de décision (Read et al., 2024).

Les tentatives de combinaison des signaux de familiarité des modèles Hebbien et anti-Hebbien reposent sur une initiative purement exploratoire. Cependant, il est plausible que les jugements de familiarité se basent sur la contribution des composantes absolue et relative de la familiarité (Bridger et al., 2014). L'analyse des performances des modèles combinés et la comparaison de celles-ci aux effets attendus du WFME (Glanzer & Bowles, 1976) permettrait d'identifier les méthodes de combinaison rendant le mieux compte des comportements humains.

Concentrons-nous sur les indices d' pour éviter les complications théoriques liées à l'interprétation du seuil en parallèle avec le criterion de la SDT. Si on s'attarde sur les effets attendus des différentes fréquences des mots sur le d', on attend de meilleures performances pour les stimuli peu fréquents (FF) (Glanzer & Adams, 1985). Ces performances meilleures se mesurent par un taux de *Hits* plus élevé et un taux de *FA* plus faible. Dans notre cas, le criterion étant intentionnellement calculé et placé pour être nul, cette différence ne peut s'expliquer que par un d' plus haut en condition FF que HF comme illustré sur la figure 14.

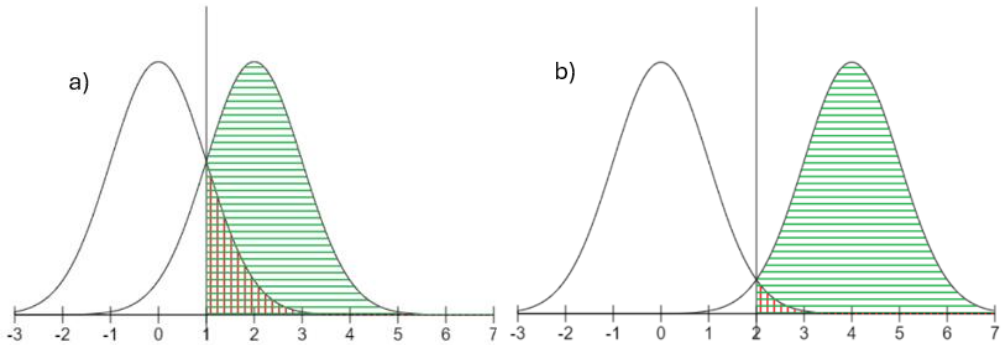

Figure 14. Distributions théoriques représentant les proportions de *Hits* en vert et de *FA* en rouge. En a), la condition théorique HF selon le WFME avec un d'=2, avec un taux de *Hits* plus faible et un taux de *FA* plus élevé que en b) représentant la condition FF pour laquelle le d' = 4.

Nos résultats, pour les méthodes par addition et multiplication, ne concorde pas avec ces effets. Le d' en condition HF est supérieur ou égal à celui en condition FF. Les possibles explications de cette discordance sont les mêmes que celles évoquées plus haut, notamment la manipulation de la familiarité absolue sans prendre en compte le caractère sémantique des stimuli.

Les méthodes envisagées dans ce travail ne semblent donc pas idéales pour rendre compte des effets des différentes conditions de familiarité absolue sur les performances humaines sur des tâches de reconnaissance. Cependant, le WFME observé chez des sujets humains ne serait pas exclusivement basé sur le principe des deux composantes de familiarité mais également sur des processus de recollection (Diana & Reder, 2006; Reder et al., 2002), avec une meilleure recollection pour les mots rares, ce qui expliquerait les meilleures performances. Cette remarque symbolise la difficulté d'interpréter les modélisations de processus cognitifs humains de manière isolée (Barsalou, 2020; Frischkorn et al., 2022).

Pour conclure, les données présentées dans ce travail n'appuient que partiellement l'hypothèse de modélisation de la familiarité absolue et relative par les modèles Hebbien et anti-Hebbien, respectivement. Les tentatives de production d'effets basés sur la distinction de ces deux composantes (i.e. WFME) ne sont pas concluantes dans ce travail. Une modélisation incluant une composante temporelle, comme envisagée par Zemliak et al. (2025), permettrait potentiellement de mettre en évidence les processus de familiarité distincts.

5.3 Limites

La condition NF, basée sur des images divisées en 25 blocs au sein desquels les pixels ont été brouillés, semblent altérer fortement les performances observées. Les modèles utilisés dans ce travail reposent fortement sur les représentations issues de ResNet50 pré-entrainé sur des millions d'images naturelles (He et al., 2016). Cette approche, bien que puissante, présente certaines limites intrinsèques dues à la nature des caractéristiques extraites. L'entrainement de ResNet50 repose sur des images bien structurées (Deng et al., 2009). Ainsi, les images présentant des structures atypiques ou artificiellement modifiées produisent des représentations potentiellement dégradées ou non informatives. Ce phénomène, connu sous le nom de décalage de domaine (« domain shift »), est une limite reconnue des modèles pré-entrainés (Torralba & Efros, 2011).

En plus de ces limites spécifiques liées à l'utilisation de ResNet50, ce travail présente plusieurs autres contraintes méthodologiques et théoriques qui doivent être mentionnées. Les modèles utilisés s'appuient sur deux règles d'apprentissage relativement simples afin d'explorer des mécanismes potentiels de reconnaissance fondés sur la familiarité. Si cette

simplicité permet une interprétation plus aisée des résultats, elle constitue également une limite en raison de la complexité réelle des mécanismes neuronaux impliqués dans la mémoire humaine (Yonelinas, 2002). Ainsi, l'absence de modélisation de processus plus sophistiqués, tels que les interactions contextuelles ou les mécanismes attentionnels (Aggleton & Brown, 2006), pourrait restreindre la portée explicative du modèle concernant le comportement humain.

Aussi, les modèles proposés traitent les stimuli de manière essentiellement statique, sans prendre en compte explicitement les dynamiques temporelles intrinsèques à la reconnaissance en mémoire. Pourtant, les processus cognitifs humains sont fondamentalement dynamiques, intégrant des informations sur la temporalité et la séquence des évènements perçus (Howard & Kahana, 2002). Un modèle intégrant une dimension temporelle, notamment à travers des architectures récurrentes ou des mécanismes prédictifs dynamiques, pourrait apporter une meilleure approximation du fonctionnement mnésique réel (Friston & Kiebel, 2009). La deuxième simulation, portant sur l'intégration pondérée des signaux Hebbien et anti-Hebbien, repose sur des méthodes simple et fixe de pondération. Cette stratégie statique ne permet pas de moduler dynamiquement la contribution de chaque signal en fonction des caractéristiques des stimuli ou du contexte expérimental comme chez certains auteurs (Tyulmankov et al., 2022). Une approche dynamique ou adaptative de la pondération, éventuellement apprise au fil des essais, pourrait mieux imiter le fonctionnement humain. De plus, la distinction artificielle des phases d'entrainement et de test par une fixation des poids limite la crédibilité de la modélisation. Une phase de test durant laquelle les poids continuent de se modifier rendrait mieux compte du fonctionnement physiologique et neuronal des traces en mémoire.

Le nombre relativement restreint d'images utilisées dans chaque condition expérimentale constitue une limite supplémentaire. Une taille d'échantillon plus importante permettrait des comparaisons aux résultats observés par Read et al. (2024).

Enfin, tout au long de ce mémoire, le terme « Hebbien » a été utilisé de façon légèrement abusive. En effet, le mécanisme nommé « Hebbien » dans ce travail ne prend pas en compte la composante temporelle qui est cruciale pour déterminer la modification de la connexion entre deux neurones (Bi & Poo, 1998). Cependant, l'utilisation du terme « Hebbien » dans ce travail s'aligne à l'usage récurrent de ce terme dans la littérature relative au sujet de la modélisation computationnelle des comportements de familiarité (Bogacz et al., 2001; Bogacz & Brown, 2003b, 2003a; Kazanovich & Borisyuk, 2021 ; Norman & O'Reilly, 2003; Tyulmankov et al., 2022).

6 Conclusion

Ce travail avait pour objectif d'étudier les mécanismes de reconnaissance en mémoire fondés sur la familiarité, en s'appuyant sur un modèle computationnel utilisé dans le travail récent de Read et al. (2024). Deux règles d'apprentissage simple (Hebbienne et anti-Hebbienne) ont visé à simuler deux formes distinctes de familiarité (absolue et relative, respectivement) dans un réseau artificiel composé d'un réseau convolutif profond pré-entrainé (ResNet50) et de deux couches entièrement connectées.

Les résultats obtenus à travers la première simulation montrent une différenciation claire entre les deux mécanismes testés. Cependant, cette différenciation ne correspond pas au WFME observé par l'expérimentation humaine. Les comportements des modèles ne semblent pas refléter fidèlement les mécanismes attendus selon la familiarité absolue ou relative. Contrairement aux hypothèses initiales, le modèle Hebbien, censé refléter la familiarité absolue, n'a pas manifesté de perturbation en condition de forte familiarité absolue. Au contraire, ses performances étaient paradoxalement meilleures en condition de haute familiarité (HF) qu'en condition de familiarité faible (FF). Ce résultat inattendu suggère que le cadre précis de notre expérimentation ne permet pas de supporter l'hypothèse d'une modélisation de la familiarité absolue par un modèle Hebbien. D'autres configurations expérimentales, avec un apprentissage répété ou un échantillon FF non sémantiquement lié à l'échantillon HF, permettraient peut-être d'éclaircir les contributions de chaque modèle au signal global de familiarité.

Le modèle anti-Hebbien, supposé représenter la familiarité relative, a démontré une robustesse à travers toutes les conditions testées, confirmant ainsi les prédictions théoriques sur sa supériorité au modèle Hebbien. Sa capacité à fournir un signal stable et discriminant, indépendamment du degré préalable de familiarité, soutient son utilité et sa robustesse dans des contextes variés mais le cadre expérimental et le manque de composante temporelle, cruciale dans la manifestation de la familiarité relative, ne permettent pas de soutenir l'hypothèse d'une modélisation de la familiarité relative par le modèle anti-Hebbien.

La deuxième simulation a envisagé de manière exploratoire les combinaisons potentielles des deux signaux (Hebbien et anti-Hebbien). Les résultats ont mis en évidence que l'addition, la multiplication ou la pondération ne permettent pas de reproduire des performances humaines observées sur des *datasets* différents selon leur niveau de familiarité absolue. Cependant, la

combinaison des signaux reste un objectif important qu'il serait possible d'explorer par d'autres techniques d'intégration, possiblement en intégrant une composante temporelle.

Ce travail présente toutefois plusieurs limites importantes. La dépendance aux caractéristiques extraites par ResNet50, la simplicité des règles neuronales utilisées, l'absence de prise en compte de la temporalité en font partie. Malgré ces limites, les résultats de ce mémoire offrent une contribution à la compréhension des mécanismes cognitifs de reconnaissance en mémoire fondés sur la familiarité. Ils soutiennent l'idée selon laquelle la reconnaissance repose sur plusieurs mécanismes interagissant de façon dynamique. Les perspectives de ce travail seraient d'inclure l'intégration d'autres mécanismes dans le processus de modélisation tels que l'interaction des types de familiarité, la temporalité ou le caractère sémantique des stimuli.

En définitive, ce travail démontre l'intérêt des approches computationnelles en neurosciences cognitives pour éclairer la complexité des mécanismes mnésiques, tout en soulignant les défis méthodologiques et théoriques qui restent à relever pour aboutir à une compréhension plus complète et intégrée de l'articulation de la mémoire humaine et plus particulièrement de la familiarité.

7 Bibliographie

Aggleton, J. P., & Brown, M. W. (2006). Interleaving brain systems for episodic and recognition memory. *Trends in Cognitive Sciences*, *10*(10), 455-463. https://doi.org/10.1016/j.tics.2006.08.003

Aggleton, J. P., McMackin, D., Carpenter, K., Hornak, J., Kapur, N., Halpin, S., Wiles, C. M., Kamel, H., Brennan, P., Carton, S., & Gaffan, D. (2000). Differential cognitive effects of colloid cysts in the third ventricle that spare or compromise the fornix. *Brain: A Journal of Neurology*, *123 (Pt 4)*, 800-815. https://doi.org/10.1093/brain/123.4.800

Anderson, N. D., Baena, E., Yang, H., & Köhler, S. (2021). Deficits in recent but not lifetime familiarity in amnestic mild cognitive impairment. *Neuropsychologia*, *151*, 107735. https://doi.org/10.1016/j.neuropsychologia.2020.107735

Anderson, N. D., Ebert, P. L., Jennings, J. M., Grady, C. L., Cabeza, R., & Graham, S. J. (2008). Recollection- and familiarity-based memory in healthy aging and amnestic mild cognitive impairment. *Neuropsychology*, *22*(2), 177-187. https://doi.org/10.1037/0894-4105.22.2.177

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In *Contemporary developments in mathematical psychology : I. Learning, memory and thinking* (p. xiii, 299-xiii, 299). W. H. Freeman.

Barsalou, L. W. (2020). Challenges and Opportunities for Grounding Cognition. *Journal of Cognition*, *3*(1). https://doi.org/10.5334/joc.116

Bastin, C., Besson, G., Simon, J., Delhaye, E., Geurten, M., Willems, S., & Salmon, E. (2019). An integrative memory model of recollection and familiarity to understand memory deficits. *The Behavioral and Brain Sciences*, *42*, e281. https://doi.org/10.1017/S0140525X19000621

Besson, G., Ceccaldi, M., & Barbeau, E. J. (2012). L'évaluation des processus de la mémoire de reconnaissance. *Revue de neuropsychologie*, *4*(4), 242-254. https://doi.org/10.1684/nrp.2012.0238

Besson, G., Ceccaldi, M., Tramoni, E., Felician, O., Didic, M., & Barbeau, E. J. (2015). Fast, but not slow, familiarity is preserved in patients with amnestic mild cognitive impairment. *Cortex*, *65*, 36-49. https://doi.org/10.1016/j.cortex.2014.10.020

Bi, G., & Poo, M. (1998). Synaptic Modifications in Cultured Hippocampal Neurons : Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *Journal of Neuroscience*, *18*(24), 10464-10472. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998

Bogacz, R., & Brown, M. W. (2003a). An anti-Hebbian model of familiarity discrimination in the perirhinal cortex. *Neurocomputing*, *52-54*, 1-6. https://doi.org/10.1016/S0925-2312(02)00738-5

Bogacz, R., & Brown, M. W. (2003b). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, *13*(4), 494-524. https://doi.org/10.1002/hipo.10093

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of Familiarity Discrimination in the Perirhinal Cortex. *Journal of Computational Neuroscience*, *10*(1), 5-23. https://doi.org/10.1023/A:1008925909305

Bowles, B., Duke, D., Rosenbaum, R. S., McRae, K., & Köhler, S. (2016). Impaired assessment of cumulative lifetime familiarity for object concepts after left anterior temporal-lobe resection that includes perirhinal cortex but spares the hippocampus. *Neuropsychologia*, *90*, 170-179. https://doi.org/10.1016/j.neuropsychologia.2016.06.035

Bridger, E. K., Bader, R., & Mecklinger, A. (2014). More ways than one : ERPs reveal multiple familiarity signals in the word frequency mirror effect. *Neuropsychologia*, *57*, 179-190. https://doi.org/10.1016/j.neuropsychologia.2014.03.007

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory : What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, *2*(1), 51-61. https://doi.org/10.1038/35049064

Brown, M. W., & Xiang, J.-Z. (1998). Recognition memory : Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology*, *55*(2), 149-189. https://doi.org/10.1016/S0301-0082(98)00002-1

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory : How the models match the data. *Psychonomic Bulletin & Review*, *3*(1), 37-60. https://doi.org/10.3758/BF03210740

Coane, J. H., Balota, D. A., Dolan, P. O., & Jacoby, L. L. (2011). Not all sources of familiarity are created equal : The case of word frequency and repetition in episodic recognition. *Memory & Cognition*, *39*(5), 791-805. https://doi.org/10.3758/s13421-010-0069-5

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(6), 923-938. https://doi.org/10.3758/bf03209340

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (s. d.). *ImageNet : A Large-Scale Hierarchical Image Database*.

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage : Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 805-815. https://doi.org/10.1037/0278-7393.32.4.805

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe : A three-component model. *Trends in Cognitive Sciences*, *11*(9), 379-386. https://doi.org/10.1016/j.tics.2007.08.001

Duke, D., Martin, C. B., Bowles, B., McRae, K., & Köhler, S. (2017). Perirhinal cortex tracks degree of recent as well as cumulative lifetime experience with object concepts. *Cortex*, *89*, 61-70. https://doi.org/10.1016/j.cortex.2017.01.015

Duzel, E., Yonelinas, A., Mangun, G., Heinze, H.-J., & Tulving, E. (1997). Event-related potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 5973-5978. https://doi.org/10.1073/pnas.94.11.5973

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*, *58-51*, ii, 32-ii, 32.

Eichenbaum, H., & Cohen, N. J. (2001). *From Conditioning to Conscious Recollection : Memory Systems of the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195178043.001.0001

Eichenbaum, H., Yonelinas, A. R., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual review of neuroscience*, *30*, 123-152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

French, R. M., Mermillod, M., Quinn, P. C., & Mareschal, D. (s. d.). *Reversing Category Exclusivities in Infant Perceptual Categorization : Simulations and Data*.

Frischkorn, G. T., Wilhelm, O., & Oberauer, K. (2022). Process-oriented intelligence research : A review from the cognitive perspective. *Intelligence*, *94*, 101681. https://doi.org/10.1016/j.intell.2022.101681

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211-1221. https://doi.org/10.1098/rstb.2008.0300

Gardette, J., Delhaye, E., & Bastin, C. (2025). The Multiple Dimensions of Familiarity : From Representations to Phenomenology. *WIREs Cognitive Science*, *16*(1), e1698. https://doi.org/10.1002/wcs.1698

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*(4), 309-313. https://doi.org/10.3758/BF03197041

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2022). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness* (arXiv:1811.12231). arXiv. https://doi.org/10.48550/arXiv.1811.12231

Gimbel, S. I., Brewer, J. B., & Maril, A. (2017). I know I've seen you before : Distinguishing recent-single-exposure-based familiarity from pre-existing familiarity. *Brain research*, *1658*, 11-24. https://doi.org/10.1016/j.brainres.2017.01.007

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8-20. https://doi.org/10.3758/bf03198438

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21-31. https://doi.org/10.1037/0278-7393.2.1.21

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (p. xi, 455). John Wiley.

Gronlund, S. D., Edwards, M. B., & Ohrt, D. D. (1997). Comparison of the retrieval of item versus spatial position information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(5), 1261-1274. https://doi.org/10.1037//0278-7393.23.5.1261

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770-778. https://doi.org/10.1109/CVPR.2016.90

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory* (p. xix, 335). Wiley.

Hendrycks, D., & Dietterich, T. (2019). *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations* (arXiv:1903.12261). arXiv. https://doi.org/10.48550/arXiv.1903.12261

Hintzman, D. L., & Caulton, D. A. (1997). Recognition memory and modality judgments : A comparison of retrieval dynamics. *Journal of Memory and Language*, *37*(1), 1-23. https://doi.org/10.1006/jmla.1997.2511

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299. https://doi.org/10.1006/jmps.2001.1388

ifigotin. (2025). imagenetmini-1000 [Dataset]. Kaggle. https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000

Jacoby, L. L. (1991). A process dissociation framework : Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541. https://doi.org/10.1016/0749-596X(91)90025-F

Jang, H., McCormack, D., & Tong, F. (2021). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLOS Biology*, *19*(12), e3001418. https://doi.org/10.1371/journal.pbio.3001418

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass : When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1534-1555. https://doi.org/10.1037/0278-7393.26.6.1534

Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics*, *10*(1), 8-14. https://doi.org/10.3758/BF03205757

Juottonen, K., Laakso, M. P., Insausti, R., Lehtovirta, M., Pitkänen, A., Partanen, K., & Soininen, H. (1998). Volumes of the Entorhinal and Perirhinal Cortices in Alzheimer's Disease. *Neurobiology of Aging*, *19*(1), 15-22. https://doi.org/10.1016/S0197-4580(98)00007-4

Kazanovich, Y., & Borisyuk, R. (2021). A computational model of familiarity detection for natural pictures, abstract images, and random patterns : Combination of deep learning and anti-Hebbian training. *Neural Networks*, *143*, 628-637. https://doi.org/10.1016/j.neunet.2021.07.022

Köhler, S., & Martin, C. B. (2020). Familiarity impairments after anterior temporal-lobe resection with hippocampal sparing : Lessons learned from case NB. *Neuropsychologia*, *138*, 107339. https://doi.org/10.1016/j.neuropsychologia.2020.107339

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway : An expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, *17*(1), 26-49. https://doi.org/10.1016/j.tics.2012.10.011

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory : A User's Guide* (2e éd.). Psychology Press. https://doi.org/10.4324/9781410611147

Mandler, G. (1980). Recognizing : The judgment of previous occurrence. *Psychological Review*, *87*(3), 252-271. https://doi.org/10.1037/0033-295X.87.3.252

Martin, C. B., Cowell, R. A., Gribble, P. L., Wright, J., & Köhler, S. (2016). Distributed category-specific recognition-memory signals in human perirhinal cortex. *Hippocampus*, *26*(4), 423-436. https://doi.org/10.1002/hipo.22531

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Scipy*. https://doi.org/10.25080/Majora-92bf1922-00a

Mecklinger, A., & Bader, R. (2020). From fluency to recognition decisions : A broader view of familiarity-based remembering. *Neuropsychologia*, *146*, 107527. https://doi.org/10.1016/j.neuropsychologia.2020.107527

Milner, B., Squire, L. R., & Kandel, E. R. (1998). Cognitive neuroscience and the study of memory. *Neuron*, *20*(3), 445-468. https://doi.org/10.1016/s0896-6273(00)80987-3

Monaco, J. D., Abbott, L. F., & Kahana, M. J. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning & Memory*, *14*(3), 204-213. https://doi.org/10.1101/lm.363207

Murray, E. A., & Bussey, T. J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Sciences*, *3*(4), 142-151. https://doi.org/10.1016/s1364-6613(99)01303-0

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Current Opinion in Neurobiology*, *11*(2), 188-193. https://doi.org/10.1016/s0959-4388(00)00195-1

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory : A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611-646. https://doi.org/10.1037/0033-295X.110.4.611

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary Learning Systems. *Cognitive Science*, *38*(6), 1229-1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065-1076. https://doi.org/10.1214/aoms/1177704472

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch : An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*.

https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Pitarque, A., Meléndez, J. C., Sales, A., Mayordomo, T., Satorres, E., Escudero, J., & Algarabel, S. (2016). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection, familiarity and false recognition, estimated by an associative process-dissociation recognition procedure. *Neuropsychologia*, *91*, 29-35. https://doi.org/10.1016/j.neuropsychologia.2016.07.010

Popov, V., & Reder, L. (s. d.). *Frequency Effects in Recognition and Recall*. Consulté 14 août 2025, à l'adresse https://academic.oup.com/edited-volume/57928/chapter/475476057

Read, J., Delhaye, E., & Sougné, J. (2024). Computational models can distinguish the contribution from different mechanisms to familiarity recognition. *Hippocampus*, *34*(1), 36-50. https://doi.org/10.1002/hipo.23588

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A Reexamination of Stimulus-Frequency Effects in Recognition : Two Mirrors for Low- and High-Frequency Pseudowords. *Journal of experimental psychology. Learning, memory, and cognition*, *28*(1), 138-152. https://doi.org/10.1037//0278-7393.28.1.138

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency : A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294-320. https://doi.org/10.1037/0278-7393.26.2.294

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Snodgrass, J. G., & Burns, P. M. (1978). The effect of repeated tests on recognition memory for pictures and words. *Bulletin of the Psychonomic Society*, *11*(4), 263-266. https://doi.org/10.3758/BF03336826

Snodgrass, J. G., & McClure, P. (1975). Storage and retrieval properties of dual codes for pictures and words in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(5), 521-529. https://doi.org/10.1037/0278-7393.1.5.521

Snodgrass, J. G., Wasser, B., Finkelstein, M., & Goldberg, L. B. (1974). On the fate of visual and verbal memory codes for pictures and words : Evidence for a dual coding mechanism in recognition memory. *Journal of Verbal Learning & Verbal Behavior*, *13*(1), 27-37. https://doi.org/10.1016/S0022-5371(74)80027-7

Squire, L. R. (2004). Memory systems of the brain : A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171-177. https://doi.org/10.1016/j.nlm.2004.06.005

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207-222. https://doi.org/10.1080/14640747308400340

Suzuki, W. A., & Naya, Y. (2014). The Perirhinal Cortex. *Annual Review of Neuroscience*, *37*(Volume 37, 2014), 39-53. https://doi.org/10.1146/annurev-neuro-071013-014207

torchvision—Torchvision 0.22 documentation. (s. d.). Consulté 4 août 2025, à l'adresse https://docs.pytorch.org/vision/stable/

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521-1528. https://doi.org/10.1109/CVPR.2011.5995347

Tsivilis, D., Vann, S. D., Denby, C., Roberts, N., Mayes, A. R., Montaldi, D., & Aggleton, J. P. (2008). A disproportionate role for the fornix and mammillary bodies in recall versus recognition memory. *Nature Neuroscience*, *11*(7), 834-842. https://doi.org/10.1038/nn.2149

Tulving, E. (1972). Episodic and semantic memory. In *Organization of memory* (p. xiii, 423-xiii, 423). Academic Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology / Psychologie canadienne*, *26*(1), 1-12. https://doi.org/10.1037/h0080017

Tyulmankov, D., Yang, G. R., & Abbott, L. F. (2022). Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron*, *110*(3), 544-557.e8. https://doi.org/10.1016/j.neuron.2021.11.009

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., …

van Mulbregt, P. (2020). SciPy 1.0 : Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Welcome to Python.org. (2025, juillet 24). Python.Org. https://www.python.org/

Wolk, D. A., Signoff, E. D., & DeKosky, S. T. (2008). Recollection and familiarity in amnestic mild cognitive impairment : A global decline in recognition memory. *Neuropsychologia*, *46*(7), 1965-1978. https://doi.org/10.1016/j.neuropsychologia.2008.01.017

Xiang, J.-Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, *37*(4-5), 657-676. https://doi.org/10.1016/S0028-3908(98)00030-6

Yang, H., McRae, K., & Köhler, S. (2023). Perirhinal cortex automatically tracks multiple types of familiarity regardless of task-relevance. *Neuropsychologia*, *187*, 108600. https://doi.org/10.1016/j.neuropsychologia.2023.108600

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., & Gilmer, J. (2020). *A Fourier Perspective on Model Robustness in Computer Vision* (arXiv:1906.08988). arXiv. https://doi.org/10.48550/arXiv.1906.08988

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory : Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341-1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (2002). The nature of recollection and familiarity : A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and Familiarity : Examining Controversial Assumptions and New Directions. *Hippocampus*, *20*(11), 1178-1194. https://doi.org/10.1002/hipo.20864

Yonelinas, A. P., Ramey, M. M., & Riddell, C. (2024). Recognition Memory : The Role of Recollection and Familiarity. In M. J. Kahana & A. D. Wagner (Éds.), *The Oxford Handbook of Human Memory, Two Volume Pack* (1re éd., p. 923-958). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190917982.013.32

Yu, J., Li, R., Jiang, Y., Broster, L. S., & Li, J. (2016). Altered Brain Activities Associated with Neural Repetition Effects in Mild Cognitive Impairment Patients. *Journal of Alzheimer's Disease*, *53*(2), 693-704. https://doi.org/10.3233/JAD-160086

Zemliak, V., Pipa, G., & Nieters, P. (2025). Continual familiarity decoding from recurrent connections in spiking networks. *PLOS Computational Biology*, *21*(8), e1013304. https://doi.org/10.1371/journal.pcbi.1013304