

Modélisation de la familiarité grâce à la combinaison d'un réseau profond avec un apprentissage Hebbien

Auteur : Read, John

Promoteur(s) : Sougne, Jacques

Faculté : Faculté de Psychologie, Logopédie et Sciences de l'Éducation

Diplôme : Master en sciences psychologiques, à finalité spécialisée en psychologie clinique

Année académique : 2021-2022

URI/URL : <http://hdl.handle.net/2268.2/16184>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



Faculté de Psychologie, Logopédie et
Sciences de l'Éducation

Modélisation de la familiarité grâce à la combinaison d'un réseau profond avec un apprentissage Hebbien

Mémoire de fin d'études en vue de l'obtention du grade de Master en Sciences
Psychologiques, à finalité spécialisée en psychologie clinique, orientation neuropsychologie

John READ

Promoteur : Dr. Jacques SOUGNÉ

Lecteurs : Dr. Christine BASTIN
Dr. Daniel DEFAYS

Année académique 2021-2022

Ce mémoire fût un travail de longue haleine ainsi qu'un véritable défi pour un novice tel que moi. Aussi, je tiens à remercier chaleureusement toutes les personnes qui ont contribué de près ou de loin à son achèvement.

Mon promoteur, Dr. Jacques Sougné, qui m'a fait confiance tout au long de ce projet et a su me guider avec bienveillance. Votre enseignement m'est très précieux et je vous remercie de m'avoir fait découvrir un domaine aussi fascinant que celui-ci.

Stephan, mon ami de longue date, qui m'a patiemment appris les rudiments du codage en Python. Tu as eu la gentillesse de m'accompagner pendant toute cette année et pour ça je te remercie. Je n'aurais jamais pu aller aussi loin sans ton aide.

Jean-Pierre et Fanny qui ont pris de leur temps pour vérifier mon orthographe parfois questionnable. Vos corrections ont grandement contribué à la qualité de ce travail.

Dr. Gregory Hammad qui m'a offert son aide lorsque j'en avais besoin. Je regrette de ne pas avoir pu travailler plus longtemps avec toi.

Dr. Roman Borisjuk qui a gentiment accepté de répondre à mes questions sur son article.

Mes lecteurs, Dr. Christine Bastin et Dr. Daniel Defays, qui ont eu l'amabilité de porter un intérêt particulier à mon travail.

Mes parents, Colin et Camille qui m'ont toujours soutenu durant mes études, dans les bons comme dans les moins bons moments. Merci de m'avoir accompagné dans mes choix et mes projets.

TABLE DES MATIERES

LISTE DES ABRÉVIATIONS	4
INTRODUCTION.....	5
PARTIE I : THÉORIE	7
1. Mémoire et reconnaissance	8
1.1. La mémoire épisodique	8
1.2. La mémoire de reconnaissance	9
2. Théories du Double-Processus	10
2.1. Modèle d'Atkinson	10
2.2. Modèle de Mandler	11
2.3. Modèle de Jacoby	12
2.4. Modèle de Tulving	13
2.5. Modèle du Dual-Process Signal Detection	14
3. Arguments en faveur des TDP	16
3.1. Vitesse de récupération	17
3.2. Receiver Operating Characteristics.....	17
3.3. Données neuro-anatomiques	19
3.4. Données électrophysiologiques	21
4. Évaluation de la familiarité	22
4.1. Tâches neuropsychologiques	22
4.1.1. Reconnaissance Oui/Non.....	22
4.1.2. Reconnaissance à choix forcés	23
4.1.3. Paradigme de mémoire source.....	23
4.2. Phénomènes comportementaux	23
4.2.1. Capacité de stockage illimitée	24
4.2.2. Effet de récence et de primauté	24
4.2.3. Effet de similarité et fausses reconnaissances	24
5. Modèles computationnels.....	25
5.1. Modèles combinés	26
5.2. Modèles spécialisés.....	29
5.2.1. Modèle anti-Hebbien	30
5.2.2. Modèle Hebbien	30
5.3. Limitation de l'activité des neurones de nouveauté.....	31
5.3.1. Compétition inhibitrice.....	32
5.3.2. Fortes connexions	32

5.4. Modèle FaRe	33
6. Objectifs du mémoire	36
PARTIE II : MODÉLISATION	38
1. Réseaux de neurones artificiels	39
2. Architecture du modèle	41
2.1. Module d'extraction des caractéristiques.....	43
2.2. Module de mémoire Hebbien.....	45
2.3. Paramètres de base du réseau.....	48
3. Simulations.....	49
3.1. Capacité de mémoire.....	49
3.1.1. Dataset	49
3.1.2. Méthodologie.....	49
3.1.3. Résultats.....	50
3.2. Effets de récence et de primauté	53
3.2.1. Méthodologie.....	53
3.2.2. Résultats.....	54
3.3. Similarité visuelle	56
3.3.1. Dataset	56
3.3.2. Méthodologie.....	56
3.3.3. Analyses statistiques.....	57
3.3.4. Résultats.....	57
3.4. Orientation de la cible	59
3.4.1. Dataset	59
3.4.2. Méthodologie.....	59
3.4.3. Résultats.....	60
PARTIE III : DISCUSSION	61
1. Capacité de stockage	63
1.1. Corrélations entre les entrées	64
1.2. Taille du module de mémoire	64
1.3. Oubli catastrophique	65
2. Rôle des effets de récence et de primauté	67
2.1. Effet de récence.....	68
2.2. Effet de primauté.....	69
3. Similarité visuelle.....	70
3.1. Erreur de familiarité et phénomène de Déjà-Vu	71

3.2. Format du test	73
4. Familiarité des visages humains.....	74
4.1. Pré-entraînement du réseau.....	75
4.2. Dataset chaotique.....	75
5. Comparaison entre les modèles.....	76
5.1. Nature de la distribution.....	76
5.2. Plasticité synaptique.....	78
5.3. Module d'extraction.....	81
6. Limites et implications	83
6.1. Simplicité du modèle	83
6.2. Apprentissage du modèle	84
6.3. Code en open access	85
CONCLUSIONS.....	86
BIBLIOGRAPHIE	89
ANNEXES	101

LISTE DES ABRÉVIATIONS

CLS Complementay Learning Systems

DPSD Dual-Process Signal Detection

DV Déjà-Vu

FaRe Familiarity Recognition model

IA Intelligence Artificielle

LTD Long-Term Depression

LTP Long-Term Potentiation

MdR Mémoire de Reconnaissance

ME Mémoire Épisodique

MS Mémoire Sémantique

RCF Reconnaissance à Choix Forcés

RNA Réseaux de Neurones Artificiels

ROC Receiver Operating Characteristics

RPC Réseau Profond Convolutif

SdF Sentiment de Familiarité

TDP Théories du Double-Processus

INTRODUCTION

Les modèles informatiques nous ont toujours aidés à appréhender certains phénomènes auparavant mystérieux, tels que l'évolution d'une étoile, les prédictions météorologiques ou encore les comportements de reproduction chez certains animaux. Dès lors, il est parfaitement légitime de se questionner sur l'apport de ces modèles d'intelligence artificielle (IA) au domaine de la psychologie.

L'IA a été définie par Shapiro (1992) comme « les domaines de la science et de l'ingénierie qui traitent de la compréhension, à l'aide de l'ordinateur, de ce qui est appelé comportement intelligent, et de la création de systèmes artificiels qui reproduisent ces comportements dits intelligents ». Ainsi, l'un des enjeux principaux de l'IA serait d'arriver à imiter l'homme et à reproduire ses compétences et capacités de manière artificielle. Ces 40 dernières années, nombreux sont les chercheurs ayant développé et testé des modèles informatiques afin d'étudier la cognition. Ceux-ci ont largement contribué à une meilleure compréhension des processus sous-jacents à la cognition humaine (voir Defays et al. (1997) pour une synthèse). En effet, la modélisation cognitive nous permet de recréer les compétences humaines, et ce, grâce à des modèles formels et explicatifs des processus cognitifs. Implémenter ces modèles nous permet ainsi de clarifier nos théories, d'en créer des nouvelles mais également de générer des prédictions et des hypothèses supplémentaires (Zuidema et al., 2020).

Parmi tous les courants ayant tenté de modéliser les comportements humains, le connexionnisme fait partie des plus importants en ce qui concerne les apports de l'IA à la psychologie. Popularisé par Rumelhart & McClelland (1986), cette approche de modélisation des phénomènes mentaux et comportementaux est grandement inspirée du fonctionnement du système nerveux. En effet, Hebb (1949) pensait que « les activités mentales ne sont que la résultante de l'activité parallèle d'unités élémentaires interconnectées », autrement dit les neurones. Le mouvement du connexionnisme conçoit l'activité mentale comme une aptitude à associer des idées entre elles. La reconnaissance d'objets ainsi que l'apprentissage des associations entre eux sont donc des éléments primordiaux pour le bon fonctionnement des modèles connexionnistes (Defays et al., 1997). Dans ce type d'architecture – appelé réseaux de neurones artificiels (RNA) –, la connaissance réside donc dans des connexions pondérées entre des unités les plus élémentaires possible qui, isolées, ne sont porteuses d'aucune signification (Mitchell, 2019). En s'inspirant du fonctionnement des neurones du cerveau, ces modèles seraient capables d'apprendre par eux même à partir de données et de leurs interactions avec le

monde dans lequel ils évoluent.

Dans le cadre de ce mémoire, nous nous intéresserons à un phénomène de la psychologie cognitive bien précis – à savoir la familiarité – que nous allons tenter de modéliser dans une perspective cognitiviste. Ce phénomène correspond au sentiment plus ou moins fort de savoir que quelque chose (une situation, un évènement, ...) a déjà été rencontré par le passé (Tulving, 1985). Sur base des précédents modèles de familiarité (Bogacz et al., 2001b ; Kazanovich & Borisjuk, 2021), nous avons implémenté un nouveau modèle connexionniste dans le langage informatique Python. Nous avons ensuite simulé une tâche de reconnaissance, durant laquelle notre modèle a dû discriminer entre une nouvelle image et une ancienne image apprise lors de la phase d’entraînement. Nous avons ainsi pu explorer une large variété de données comportementales.

Grâce à ce mémoire, nous espérons apporter de nouvelles informations sur les mécanismes mis en œuvre lors d’une situation faisant appel à un jugement de reconnaissance, et plus précisément à la familiarité. Aussi, ce travail pourrait apporter des preuves supplémentaires en faveur des théories et autres modèles de la reconnaissance, ainsi que des circuits neuronaux impliqués lors d’une décision de reconnaissance basée sur la familiarité. En effet, la majeure partie des travaux en modélisation cognitive s’inspire directement des modèles biologiques du cerveau en ce qui concerne la conception des réseaux. L’architecture multicouche du cortex visuel est, par exemple, largement utilisée pour la reconnaissance d’images (le Cun, 2019).

Ce mémoire est scindé en trois parties. La première partie de ce travail est consacrée à l’exploration théorique du processus de familiarité. Nous nous attardons également sur les précédentes tentatives de modélisation du phénomène. Dans la seconde partie, après avoir détaillé l’architecture de notre réseau ainsi que la méthodologie des simulations réalisées, nous présentons les résultats de nos différentes modélisations. Ces résultats sont ensuite discutés dans la troisième partie de ce mémoire. Nous esquissons des pistes d’améliorations et de futures directions. En parallèle, nous épinglons certaines limites de notre travail ainsi que du domaine de l’IA en général.

PARTIE I :

THÉORIE

Avant d'envisager la modélisation, il convient d'exposer certains points théoriques qui nous permettront de comprendre le fonctionnement de la reconnaissance chez l'homme. Plus précisément, nous distinguerons la recollection de la familiarité. Nous passerons en revue les différentes théories explicatives et les arguments en faveur de cette distinction. Enfin, nous décrirons les précédents modèles computationnels ayant trait à la familiarité.

1. Mémoire et reconnaissance

La mémoire est probablement l'un des phénomènes les plus complexes des neurosciences cognitives. Au fil du temps, de nombreuses taxonomies de la mémoire ont émergé, avec chacune leurs divergences sur le plan fonctionnel mais également anatomique (Squire & Zola, 1998 ; Tulving, 1983). Situons d'abord la familiarité dans un cadre théorique clair et précis, en accord avec les découvertes récentes.

1.1. La mémoire épisodique

Parmi les différents types de mémoires documentés, c'est la mémoire épisodique (ME) qui correspond le mieux au terme commun « mémoire ». Elle désigne la capacité à se souvenir d'objets ou de personnes qui ont été rencontrés par le passé mais également d'événements spécifiques qui ont été personnellement vécus (Tulving, 1983). La ME permet de situer un épisode de vie dans son contexte d'apprentissage, c'est-à-dire dans son contexte spatio-temporel. Selon Squire et coll. (1998, 2004), elle fait partie, avec la mémoire sémantique (MS)¹, de la mémoire déclarative. L'intégrité de la ME dépendrait entre autres du lobe temporal médial ainsi que des régions frontales du cortex cérébral (Squire, 2004).

Le souvenir d'un événement serait sous-tendu par deux mécanismes successifs : l'encodage et la récupération (Tulving, 1983). L'encodage est le processus via lequel les caractéristiques d'un stimulus ainsi que son contexte spatio-temporel sont traités puis convertis en une trace mnésique. Ultérieurement, lorsque des indices sont perçus dans un environnement, les mécanismes de récupération s'activent afin de récupérer la trace précédemment stockée.

Pour plusieurs chercheurs (McClelland et al., 1995 ; Schacter et al., 1998), une expérience vécue est caractérisée par un ensemble de traits (i.e. *patterns*) éparpillés au sein du cerveau. Dans cette perspective dite constructiviste, la récupération d'un souvenir nécessite dès lors un mécanisme permettant la reconstruction de ces *patterns* (McClelland et al., 1995). Lorsque qu'un indice de récupération est perçu dans un environnement, certains *patterns* seront

¹ La MS correspond à la mémoire de nos connaissances sur le monde (Tulving, 1983).

ainsi réactivés. L'activation va ensuite se propager au reste des caractéristiques du souvenir, permettant finalement sa recollection.

Deux processus de récupération permettraient cette complétion de *patterns* (Schacter et al., 1998). Premièrement, les processus stratégiques, ou processus de rappel, qui correspondent à une recherche active en mémoire et nécessitent un effort cognitif. Deuxièmement, les processus associatifs, qui, lorsque le chevauchement entre la trace mnésique et l'indice de récupération est suffisant, vont activer automatiquement cette trace et l'envoyer vers la conscience (Tulving, 1985). Ces processus associatifs, dont fait partie la familiarité, font référence à la mémoire de reconnaissance, qui fait l'objet de la sous-section suivante.

1.2. La mémoire de reconnaissance

La mémoire de reconnaissance (MdR), ou plus simplement la reconnaissance, permet donc l'activation automatique d'une trace en mémoire, en la rendant accessible à la conscience (Tulving, 1985). Plus précisément, elle correspond à notre capacité à juger si nous avons déjà été confronté par le passé ou non à un stimulus particulier qui nous est présenté (Besson et al., 2012 ; Mandler, 1980). En comparaison avec d'autres types de mémoires, la MdR ne nécessite pas systématiquement le rappel des détails précis liés à l'évènement afin de *reconnaître* une personne, un endroit ou objet (Mandler, 1980). En effet, elle se baserait sur la perception d'un stimulus externe présent dans l'environnement et ne dépendrait donc pas d'un processus mental (Besson et al., 2012). Toutefois, dans la majorité des cas, ces détails seront également acquis en cours du processus mnésique, que ce soit grâce aux indices présents dans l'environnement ou aux processus stratégiques. Ainsi, la MdR se positionne à l'interface entre les fonctions mnésiques et les fonctions perceptives (Besson et al., 2012).

Pour illustrer ces propos, voici un exemple inspiré du « boucher dans le bus » (i.e. *butcher-on-the-bus*) proposé par Mandler (1980). Dans cette situation, je croise une personne en allant faire mes courses un samedi au magasin du coin. Dans un premier temps, je reconnais l'individu comme étant quelqu'un de familier, mais sans être capable de l'identifier ni de déterminer la date ou le lieu de notre précédente rencontre. Je fouille ensuite dans ma mémoire afin de déterminer de qui il s'agit. Les processus stratégiques entrent en jeu et des questions émergent : « Où ai-je vu cette personne ? À l'université ? Dans le bus ? ». A un moment, cela me revient, il s'agit du chauffeur du bus 48, que j'ai croisé hier en allant à l'université !

L'exemple proposé par Mandler est propre à sa conception de la MdR, qui diffère sur certains points d'autres modèles théoriques (voir infra). Néanmoins, il nous permet

d'appréhender la reconnaissance comme étant double. En effet, avant les années 70, la MdR était surtout envisagée sous l'angle d'un processus unique de reconnaissance (Egan, 1958). Selon les auteurs partisans de cette théorie, la MdR est représentée par un continuum sur lequel la force d'une trace mnésique varie en intensité, allant de faible jusqu'à forte (Squire et al., 2007 ; Squire & Wixted, 2011). Cependant, les conceptions actuelles – appelées théories du double-processus (TDP) – envisagent plutôt la reconnaissance selon deux processus distincts mais complémentaires : la *recollection* et la *familiarité*.

Plus concrètement, la familiarité correspondrait à un « sentiment de vieillesse » à propos d'un stimulus ; elle donnerait ainsi des indications quant au fait que cet événement a déjà été vécu auparavant (Yonelinas et al., 2010). En d'autres termes, le sentiment de familiarité (SdF) serait une mesure quantitative d'un souvenir, permettant de distinguer l'ancien du nouveau sur base des caractéristiques perceptives et/ou des informations spécifiques à une situation vécue précédemment (Tulving, 1985 ; Yonelinas & Jacoby, 1994). Ainsi, dans un type de reconnaissance basé sur la familiarité, l'individu sait qu'il a déjà vécu une expérience par le passé mais n'est pas capable d'en identifier la source en récupérant les détails relatifs à l'encodage (Cleary, 2008). A l'inverse, la recollection correspond à la faculté de se souvenir explicitement d'un événement vécu, ainsi que de le revivre mentalement (Tulving, 1985 ; Yonelinas et al., 2010). Dans un type de reconnaissance basée sur la recollection, l'individu se rappelle les détails qualitatifs liés à l'encodage, tel que le contexte spatio-temporel, d'une situation passée (Yonelinas & Jacoby, 1994). Ces deux mécanismes, illustrés dans l'exemple ci-dessus, permettent à eux deux de discerner les stimuli déjà rencontrés des nouvelles situations.

2. Théories du Double-Processus

Les sous-sections suivantes retraceront la progression scientifique des TDP afin d'éclaircir le fonctionnement des mécanismes de familiarité et de recollection (voir Yonelinas (2002) pour une revue extensive de la littérature à ce sujet). Nous approfondirons davantage le mécanisme de familiarité qui fait l'objet de ce mémoire.

2.1. Modèle d'Atkinson

Dans le modèle proposé par Atkinson et coll. (Juola et al., 1971), la reconnaissance d'un stimulus s'exprime sous la forme d'une réponse rapide basée sur la familiarité de ce dernier. Dans le cas où la réponse est perçue comme ambiguë, un processus de recherche active en mémoire s'enclenche. Lorsqu'un stimulus est visualisé, sa trace mnésique augmente

temporairement. Cette augmentation le distinguera le cas échéant d'un leurre, c'est-à-dire un stimulus qui n'a pas été vu. Ainsi, lors d'un test de reconnaissance, un stimulus étudié, c'est-à-dire une cible, sera perçu comme plus familier qu'un leurre.

Dans leur modèle, les auteurs postulent qu'il existerait un critère de décision au-delà duquel un item serait considéré comme ancien, ainsi qu'un second critère en deçà duquel un item serait considéré comme nouveau. La distribution théorique de l'activation des cibles et des leures correspond à deux courbes Gaussiennes qui se chevauchent partiellement (**Figure 1**). La recollection s'enclenche lorsque l'activation d'un item est située entre les deux critères, ce qui se traduirait par un délai de réponse supérieur.

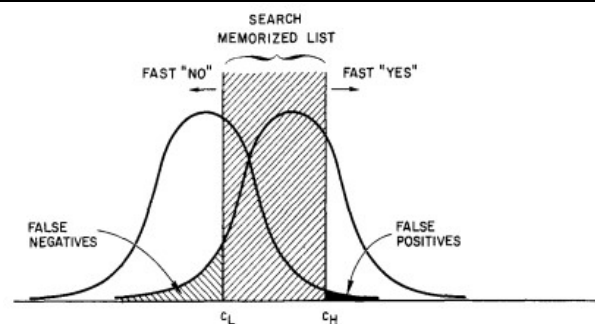


Figure 1 | Distribution de l'activation (axe des abscisses) des items cibles (courbe de droite) et des leures (courbe de gauche) selon le modèle d'Atkinson (Juola et al., 1971).

2.2. Modèle de Mandler

Dans le modèle proposé par Mandler (1980), la reconnaissance peut s'effectuer via un processus de familiarité, par lequel un item est jugé comme ancien s'il atteint un seuil, similairement au modèle d'Atkinson. Dans cette conception, l'augmentation de la familiarité d'un stimulus correspond à l'intégration des caractéristiques perceptives au sein même du stimulus. Cette reconnaissance peut également s'effectuer de façon parallèle via un processus de recollection, lequel consiste en une recherche des informations inter-items, c'est-à-dire des détails liés au contexte d'encodage, au sein de la mémoire à long terme. Ainsi, dans le modèle de Mandler, les deux processus sont considérés comme indépendants mais agissant en parallèle. La reconnaissance par la familiarité serait aussi plus rapide que la recollection.

L'exemple cité précédemment, adapté du phénomène du *Butcher-on-the-bus* proposé par Mandler (1980) distingue ces deux processus. Au cœur de cette distinction se trouve l'idée que la familiarité correspond à une incapacité à se souvenir du contexte de l'épisode antérieur et de toute information associative qui en expliquerait l'origine.

2.3. Modèle de Jacoby

Dans le modèle décrit par Jacoby et coll. (Jacoby, 1991 ; Jacoby & Dallas, 1981), la recollection et la familiarité sont bien indépendantes et agissent en parallèle. La première serait un processus de rappel contrôlé se basant sur l'élaboration². La seconde serait quant à elle un processus automatique lié à la fluence perceptive d'un stimulus. Ici aussi, la reconnaissance par la familiarité serait plus rapide que la recollection.

La *fluence perceptive* d'un stimulus fait référence au SdF qui apparaît lorsqu'un individu considère qu'une amélioration de la vitesse et de la facilité avec laquelle un stimulus est traité, autrement dit sa fluidité de son traitement, est un signe que ce stimulus a déjà été rencontré. En effet, Jacoby & Dallas (1981) ont montré que quand une information est vécue à nouveau, elle est traitée plus facilement et plus rapidement que lors de la première fois. A force des répétitions, une amélioration de la fluence perceptive donnerait naissance au sentiment de « vieillesse » par rapport à un stimulus. Ceci amènerait dès lors un individu à considérer cette situation comme ayant déjà été vécue (Geurten et al., 2020 ; Jacoby & Whitehouse, 1989). Selon Jacoby & Whitehouse (1989), un SdF peut survenir face à de nouveaux stimuli via la manipulation de la fluidité de traitement de ceux-ci. La manipulation de cette fluence perceptive aura par ailleurs un impact significatif sur le processus de familiarité, mais n'impactera pas le processus de recollection.

En outre, Jacoby (1991) postule que le SdF s'exprime également à travers une amélioration de la *fluence conceptuelle* d'un stimulus, c'est-à-dire une amélioration de la fluidité du traitement de sa signification. Par exemple, lorsqu'un contexte sémantiquement proche est attribué à un stimulus, ce stimulus sera plus facilement considéré comme étant familier (Whittlesea, 1993). Des études (Wolk et al., 2005, 2009) suggèrent effectivement que la manipulation de la fluence conceptuelle d'un stimulus impacte significativement le SdF.

En conséquence, la familiarité n'est pas supposée refléter le fonctionnement d'un système mnésique différent de la recollection, pas plus qu'elle ne s'appuie exclusivement sur l'activation de représentations déjà existantes (Jacoby, 1991 ; Jacoby & Dallas, 1981). Au contraire, le modèle de Jacoby suggère que la familiarité, à l'instar de la recollection, repose sur un souvenir détaillé d'une expérience précédente.

² L'élaboration correspondrait à la création de liens fonctionnels entre un événement et son contexte. Ce point ne sera pas plus développé.

Focus. Procédure de Dissociation des Processus

Jacoby et coll. (Jacoby, 1991 ; Jacoby et al., 1993) ont proposé le paradigme de la procédure de dissociation des processus en se basant sur le postulat que si un sujet parvient à recollecter un item, il devrait alors être capable de déterminer soit quand, soit où il a été initialement encodé ; la familiarité ne supporterait pas cette supposition. Si recollection et familiarité sont indépendantes, il devrait ainsi être possible d'effectuer des manipulations capables d'influencer l'un sans influencer l'autre et inversement.

Illustrons avec une épreuve de complétion de mots. Le sujet se verra présenter une succession de mots. A un certain moment, un mot présenté sera incomplet et le sujet devra compléter les trois dernières lettres manquantes. Toutefois, la manière dont il doit compléter le radical changera en fonction de la consigne qui lui sera donnée. Dans la consigne d'*inclusion*, il est demandé au patient de compléter le radical pour former un mot vu précédemment ou le premier mot qui lui passe par l'esprit. Dans la consigne d'*exclusion*, le sujet doit compléter le radical pour former un mot qu'il n'a pas vu au préalable. Ce procédé permet dès lors la comparaison des performances du sujet selon la condition, à savoir d'inclusion ou d'exclusion.

Une fois l'épreuve terminée, des équations mathématiques basiques nous permettent de quantifier les processus survenus lors de la tâche. Grâce à ces équations, nous obtenons ainsi des probabilités concernant l'implication des deux processus. Nous pouvons ainsi déterminer si le sujet s'est plutôt basé sur un SdF pour répondre ou s'il a plutôt produit sa réponse grâce à une recollection.

2.4. Modèle de Tulving

Le modèle proposé par Tulving (Tulving, 1985) se rapproche fortement de celui d'Atkinson, en postulant que les performances en MdR dépendent de deux systèmes mnésiques distincts ; tous deux appartenant à la mémoire déclarative – ou mémoire explicite – qui concerne les informations récupérées consciemment (Squire, 2004). D'après l'auteur, ce serait la ME qui donnerait naissance à l'expérience consciente de *remembering*, autrement dit rappel ou recollection, tandis que ce serait la MS qui donnerait naissance au sentiment de *knowing*, c'est-à-dire à un SdF en l'absence de *remembering*³.

³ Termes volontairement laissés en anglais dans la mesure où leur traduction pourrait porter à confusion.

Dans cette conception, ces concepts correspondent respectivement à la conscience auto-noétique, c'est-à-dire le vécu subjectif que l'on ressent lors de la reviviscence d'un souvenir épisodique, ainsi qu'à la conscience noétique, qui correspond à l'état de conscience lorsque l'on pense à quelque chose de connu (Tulving, 1985). La reconnaissance en tant que telle serait située au carrefour de ces mémoires sémantique et épisodique ; les informations pourraient ainsi être récupérées séparément à partir de l'un ou l'autre des systèmes (Tulving, 1985 ; Tulving & Markowitsch, 1998).

Focus. La paradigme *Remember/Know/Guess*

Le paradigme *Remember/Know* a été développé par Tulving (1985) dans le but de mesurer la contribution de ces différents types de mémoires aux performances mnésiques générales. Il repose sur le rapport subjectif de l'état de conscience du sujet lors de sa reconnaissance. Ainsi, si la reconnaissance d'un item est accompagnée d'un souvenir d'éléments du contexte d'encodage, c'est-à-dire basée sur la recollection. Ceci correspond à une réponse de type « Je me rappelle » ou *Remember*. En revanche, si la reconnaissance est basée sur un SdF, cela correspond à une réponse de type « Je sais » ou *Know*. Ce paradigme demande donc au sujet de rapporter son expérience subjective qui accompagne la reconnaissance d'un item, dans le but d'identifier une éventuelle recollection.

Plus concrètement, lors d'une tâche de reconnaissance, si le sujet répond « oui » à un item, il devra indiquer une réponse « R » si sa reconnaissance a été accompagnée de la recollection d'un ou plusieurs éléments associés au contexte d'encodage. Si aucun élément du contexte d'encodage n'a été récupéré mais qu'il sait qu'il a rencontré l'item, sur la base d'un SdF il devra indiquer une réponse « K ». Signalons également qu'une proposition de réponse « G », pour *Guess*, a été ajoutée pour les réponses « oui » qui ont été déterminées de façon incertaine. Cela permet d'éviter d'obtenir des réponses « K » qui ne se sont pas basées sur un SdF (Besson et al., 2012 ; Gardiner, 1988).

2.5. Modèle du Dual-Process Signal Detection

Parmi les nombreux modèles conceptuels développés à partir de la TDP, il en est un particulièrement utilisé lors d'études comportementales sur la reconnaissance ; il s'agit du modèle du *Dual-Process Signal Detection* (DPSD) proposé par Yonelinas (1994, 2002). Un des avantages du modèle DPSD est de fournir un cadre théorique permettant l'intégration et la compréhension des résultats provenant d'une grande variété de paradigmes encore utilisés aujourd'hui lors de l'évaluation clinique de la MdR (Yonelinas et al., 2010).

Cette théorie postule que la familiarité et la recollection diffèrent sur deux aspects primordiaux (Yonelinas, 1994). Premièrement, lors de la reconnaissance, les deux processus ne fournissent pas le même type d'informations. Deuxièmement, ils présentent des degrés de confiance différents vis-à-vis du stimulus qui a été reconnu. Ainsi, lors de la récupération d'un souvenir en mémoire déclarative, ces deux processus sont initiés en parallèle mais seraient bel et bien indépendants l'un de l'autre. Par ailleurs, ce modèle assume que le processus de familiarité est plus rapide que le processus de recollection (Yonelinas & Jacoby, 1994).

Ainsi, la familiarité reflète un processus de détection de signaux (Egan, 1958), décrit comme le degré de correspondance entre les caractéristiques d'un stimulus perçu et celles d'un stimulus antérieur stocké en mémoire. Lorsqu'un item est perçu, il présente un indice de familiarité, qui se distribuerait de façon continue pour les différents items, sous forme d'une courbe Gaussienne (**Figure 2**). Plus cet indice est élevé et plus un individu présente un haut degré de confiance quant au fait que l'item a déjà été étudié auparavant. A l'inverse, plus cet indice est faible et plus l'individu éprouverait un haut degré de confiance quant au fait que l'item n'a jamais été vu. Un jugement de confiance incertain serait quant à lui associé à des valeurs intermédiaires pour l'indice de familiarité, représenté sur la figure par le haut de la cloche.

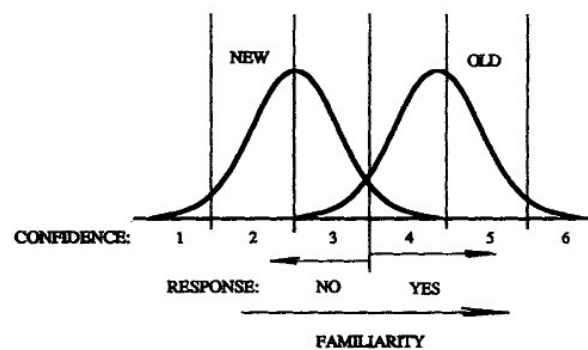


Figure 2 | Distribution des valeurs de familiarité (axe des abscisses) pour les nouveaux et anciens items selon un processus de détection de signaux (Yonelinas, 1994).

Dès lors, un haut degré de chevauchement entre les caractéristiques d'un item perçu et d'un item stocké produira un signal de familiarité relativement fort tandis qu'un faible chevauchement produira un signal plus faible (Cleary, 2008). En d'autres termes, la familiarité refléterait la « mesure quantitative de la force d'un stimulus en mémoire » (Yonelinas, 1994). Cette conception permet ainsi de rendre compte du phénomène de fausse reconnaissance, pouvant survenir lorsqu'un leurre ressemble trop fortement à une cible et déclenche dès lors un

haut niveau de familiarité (Holdstock et al., 2002 ; Norman & O'Reilly, 2003).

Par ailleurs, la recollection serait nécessaire afin de former de nouvelles associations entre objets ; la familiarité serait, quant à elle, suffisante pour la reconnaissance et ne permettrait pas les associations entre de nouveaux items (Aggleton & Brown, 1999). Signalons toutefois que si un stimulus correspond généralement à une seule pièce d'information, par exemple un objet dans une pièce (Ranganath, 2010), des associations entre items peuvent également être reconnues par familiarité (Yonelinas et al., 1999). Par exemple, l'association entre les mots « souris » et « chauve ». Néanmoins, pour que cela se produise, il faut que ces items, c'est-à-dire les différents aspects de l'évènement, soient traités comme un tout et non pris de manière isolée, selon un processus appelé *unitization* (Diana et al., 2008 ; Mayes et al., 2002 ; Quamme et al., 2007). Dans notre exemple, cette *unitization* entre les deux mots se ferait via le mot « chauve-souris », qui rassemble deux mots distincts en une seule association. Ce processus intervient notamment pour les visages, qui sont constitués des plusieurs traits tels que la bouche, le nez, les yeux, etc.

A l'opposé, la recollection serait un processus de récupération de type « tout ou rien », c'est-à-dire qu'elle a lieu ou pas (Besson et al., 2012 ; Yonelinas, 1994, 2002). Elle serait caractérisée par un seuil au-delà duquel l'individu parviendrait à recollecter les détails qualitatifs d'un évènement antérieur. En outre, en situation de recollection, la réponse donnée sera indépendante du signal de familiarité. Le degré de confiance sera très élevé étant donné la certitude d'avoir vécu l'évènement. Ce type de reconnaissance est indépendant de la familiarité et se basera uniquement sur la recollection. Dans le cas où le seuil n'est pas atteint, l'individu ne sera pas capable de se souvenir précisément des informations qualitatives liées à l'évènement en question. Une reconnaissance peut néanmoins avoir lieu sur base de la familiarité de l'item.

3. Arguments en faveur des TDP

Plusieurs auteurs mettent en évidence des distinctions empiriques entre familiarité et recollection (Diana et al., 2006 ; Yonelinas, 2002). Ces arguments confirment le postulat d'une reconnaissance qui implique plusieurs types de mémoires. Dans cette section, nous approfondirons certains de ces arguments et mettrons en évidence des propriétés de la familiarité qui seront explorées dans la suite de ce mémoire. Nous précisons que ces deux processus distincts agiraient néanmoins de manière conjointe lorsqu'une personne est face à une décision de reconnaissance ; ils seraient donc complémentaires (Gardiner, 1988). En outre, comme nous l'exposerons dans la section suivante, les tâches de reconnaissance traditionnelles

nous permettent difficilement de distinguer les rôles respectifs de ces processus lors de la récupération d'une information (Besson et al., 2012). L'importance des arguments listés ci-dessous réside dès lors dans le fait qu'ils appuient la possibilité de modéliser le SdF séparément de la recollection.

3.1. Vitesse de récupération

Une première constatation concerne la vitesse de récupération d'une information, plus rapide via la familiarité que via la recollection (Hintzman et al., 1998). En effet, les auteurs ont montré que lorsque les participants doivent répondre rapidement, ils produisent correctement des jugements de discrimination basés sur la familiarité. Cependant, ils peinent à produire ceux basés sur la recollection, laquelle nécessite le rappel des informations précises liées à l'événement. Par ailleurs, certains nouveaux items sont incorrectement considérés comme étant familiers, phénomène qui disparaît lorsque les participants disposent de plus de temps pour recollecter l'item (Rotello & Heit, 2000). Selon certaines études physiologiques (Hintzman et al., 1998 ; Xiang & Brown, 1998), une réponse de reconnaissance basée directement sur le processus de familiarité surgirait dans les 100 ms qui suivent la présentation visuelle d'un stimulus ayant déjà été étudié auparavant.

3.2. Receiver Operating Characteristics

Une seconde distinction s'observe au travers des courbes des caractéristiques des performances, ou encore courbes sensibilité/spécificité, plus fréquemment désignées par les termes anglais *Receiver Operating Characteristics* (ROC). Elles correspondent à une fonction mathématique représentant la relation entre la proportion de réponses correctes (*hits*) et de fausses alarmes et nous informe donc de la relation entre la sensibilité et la spécificité des réponses d'un participant lors d'une tâche de reconnaissance (Yonelinas & Parks, 2007). L'analyse des degrés de confiance associés aux réponses à une tâche de reconnaissance montre des courbes ROC distinctes pour la familiarité et la recollection. La familiarité serait caractérisée par une grande dispersion des degrés de confiance qu'un sujet associerait à sa réponse lors d'une tâche de reconnaissance. À l'inverse, la recollection serait caractérisée par un haut niveau de confiance. En examinant ces courbes, il serait possible d'estimer leur contribution respective lors d'une tâche de reconnaissance (Yonelinas, 1994).

Dans un paradigme ROC, le sujet doit indiquer pour chaque réponse avec quel indice de confiance, allant de 0 à 6 par exemple, il reconnaît ou non les items qui lui sont présentés (Yonelinas, 1994 ; Yonelinas & Parks, 2007). Les valeurs extrêmes correspondent aux

jugements effectués avec la plus grande certitude tandis que les valeurs intermédiaires correspondent aux jugements effectués de manière incertaine. Que ce soit pour les cibles ou les leurres, cet indice de confiance nous permet d’obtenir une mesure des distributions des réponses sur un axe de familiarité, représenté par deux courbes de Gauss de même variance (voir **Figure 2**). La courbe ROC correspond en réalité à la « fusion » de ces deux courbes Gaussiennes et représente la distribution cumulée de la proportion de réponses aux cibles en fonction de la distribution cumulée de la proportion de réponses aux distracteurs (Besson et al., 2012). Elle se trace dans l’espace $(0 ; 1) \times (0 ; 1)$ et est représentée par six points. La fonction est cumulative et chacun des points considère l’inclusion supplémentaire du pourcentage de réponses données selon l’indice de confiance suivant (Yonelinas & Parks, 2007). Pour illustrer, le premier point a comme abscisse la proportion de distracteurs auxquels le sujet a répondu avec un indice de confiance 1 et pour ordonnée la proportion de cibles auxquels il a donné un indice 1. Le deuxième point a quant à lui pour abscisse la proportion de distracteurs auxquels un indice allant de 1 à 2 a été donné et pour ordonnée la proportion de cibles avec un indice compris entre 1 et 2. Ce procédé est ainsi répété jusqu’au sixième point, qui a finalement pour abscisse la proportion de distracteurs auxquels les indices 1 à 6 ont été donnés et pour ordonnée la proportion de cibles avec les indices allant de 1 à 6 (Besson et al., 2012).

Dans les graphiques ci-dessous (**Figure 3**), deux tracés sont intéressants à analyser. Commençons par la **Figure 3A**, qui correspond à la courbe obtenue si un individu se base uniquement sur la familiarité des items pour répondre. Le premier tracé, la diagonale, relie le coin inférieur gauche au coin supérieur droit et correspond aux situations dans lesquelles le participant n’a pas réussi à discriminer correctement tous les items qui lui ont été présentés. Le second, curvilinéaire et symétrique, correspond à une discrimination parfaite entre les cibles et les distracteurs, du point de vue des indices de confiance rapportés. La **Figure 3B** nous renseigne quant à elle sur l’apparence qu’aurait la courbe ROC si l’individu venait à répondre en se basant uniquement sur le processus de recollection. On remarque qu’il s’agit d’un tracé en forme de crosse de hockey dont la majeure partie est linéaire, reflet du processus seuil caractérisant la recollection dans la DPSD. Il semble en revanche peu probable qu’un individu se base uniquement sur un SdF – ou une recollection – pour produire ses réponses (Yonelinas et al., 2010).

Dans une situation réelle, l’individu se baserait plutôt sur un mélange de recollection et de familiarité, ce qui produirait alors une courbe telle que dessinée dans la **Figure 3C**. La familiarité donne une forme en U inversé à la courbe, tandis que la recollection pousse cette

courbe vers le haut, de sorte qu'elle intersecte l'axe des ordonnées. Cette contribution donne à la courbe ROC une asymétrie par rapport à la diagonale. Ainsi, pour estimer un indice de familiarité, il faudrait calculer le degré de curvilinearité de la courbe. L'indice de recollection serait quant à lui estimé à partir de l'intersection avec l'axe des ordonnées.

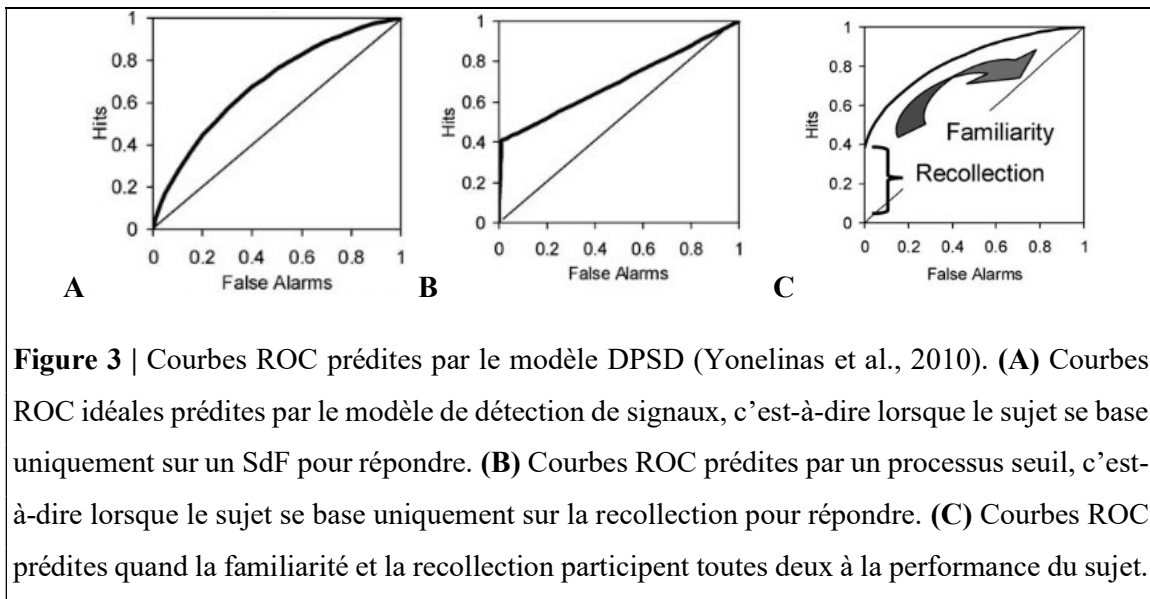


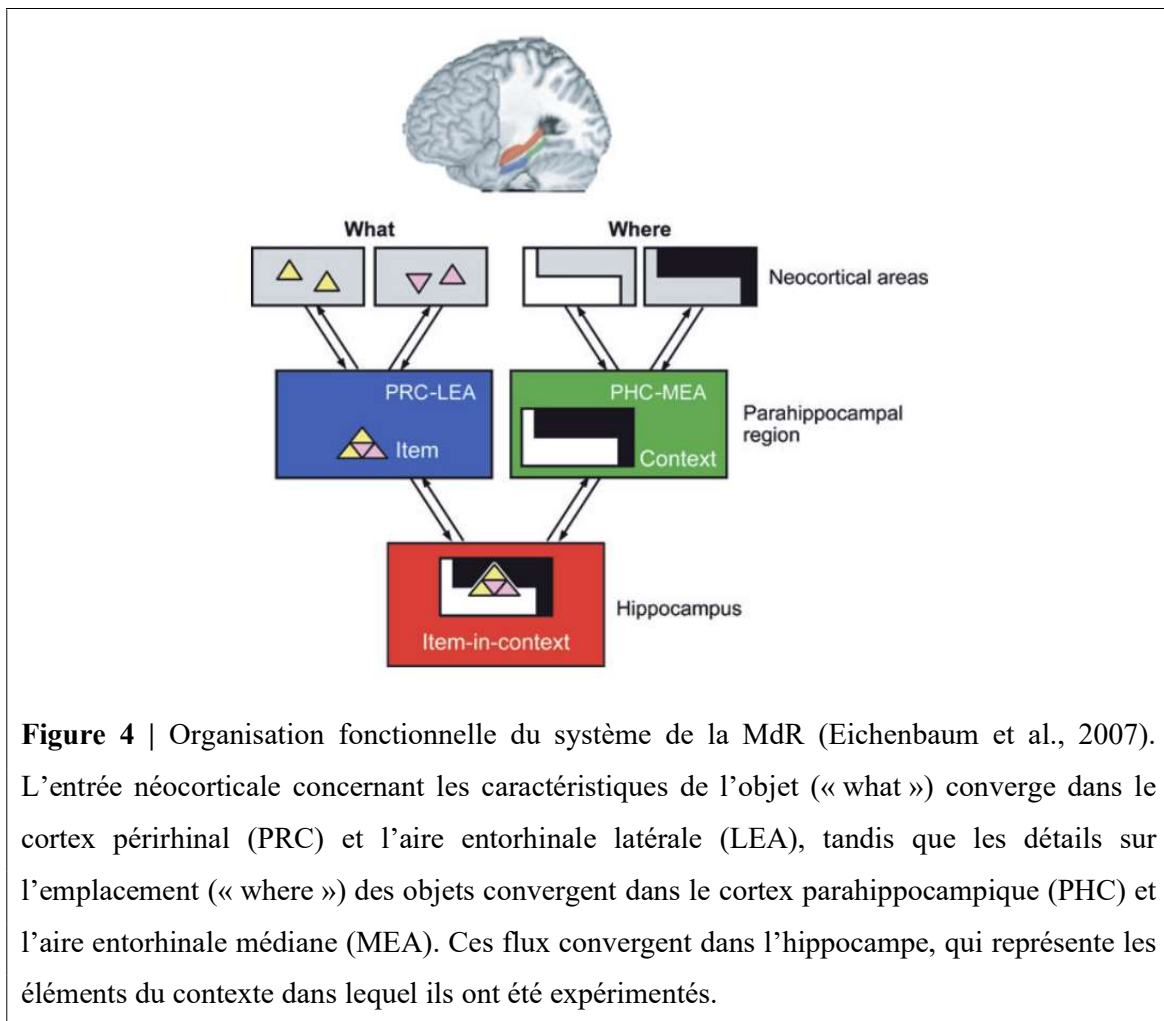
Figure 3 | Courbes ROC prédites par le modèle DPSD (Yonelinas et al., 2010). **(A)** Courbes ROC idéales prédites par le modèle de détection de signaux, c'est-à-dire lorsque le sujet se base uniquement sur un SdF pour répondre. **(B)** Courbes ROC prédites par un processus seuil, c'est-à-dire lorsque le sujet se base uniquement sur la recollection pour répondre. **(C)** Courbes ROC prédites quand la familiarité et la recollection participent toutes deux à la performance du sujet.

3.3. Données neuro-anatomiques

La distinction entre les processus de recollection et de familiarité est également confortée par une étude en neuro-imagerie réalisée avec des patients amnésiques présentant une déconnexion bilatérale du fornix (Aggleton et al., 2000). Principale voie efférente entre l'hippocampe et le diencephale, le fornix serait effectivement un élément clé du système sous-tendant la ME (Aggleton & Saunders, 1997). Les auteurs ont ainsi montré qu'une perte mnésique sévère peut survenir chez ces sujets. Cependant, des performances relativement intactes lors d'une tâche de reconnaissance chez ces mêmes patients suggèrent que certaines atteintes cérébrales n'impacteraient pas la recollection et la familiarité de la même manière. Les auteurs postulent donc l'existence d'une dissociation anatomique entre les deux processus.

Plusieurs études de neuro-imagerie suggéraient déjà une dissociation anatomique de la MdR par rapport à d'autres types de mémoires en l'associant aux régions temporales médiales du cerveau (voir Squire (2004) pour une revue à ce sujet). Des patients avec des lésions situées dans le lobe temporal médian – englobant l'hippocampe ainsi que les structures corticales alentour – présentent des déficits lors de tâches de rappel libre et de reconnaissance. Ils gardent néanmoins des performances normales lors de tâches impliquant d'autres types de mémoires non-déclaratives comme la mémoire procédurale.

Aggleton & Brown (1999) ont tenté de clarifier la contribution de l'hippocampe ainsi que des cortex du lobe temporal médian qui l'englobent – à savoir les cortex entorhinal, périrhinal et parahippocampique – à la MdR. Selon eux, l'hippocampe sous-tendrait la recollection tandis que les cortex du lobe temporal médian sous-tendraient la familiarité. Wolk et al. (2011) ont quant à eux observé une double dissociation anatomique et fonctionnelle entre familiarité et recollection en utilisant un paradigme de mémoire source (voir infra) chez des sujets âgés sains, présentant un trouble cognitif léger ou atteints de la maladie d'Alzheimer. Ils ont constaté que la recollection était plus fortement liée au volume hippocampique, tandis que la familiarité au volume des cortex entorhinal et périrhinal. En outre, de nombreux auteurs (Aggleton et al., 2005 ; Bowles et al., 2010 ; Eichenbaum et al., 2007 ; Montaldi & Mayes, 2010) ont également apporté des preuves supplémentaires concernant l'implication particulière du cortex périrhinal dans le processus de familiarité (**Figure 4**). Ainsi, des lésions sélectives du cortex périrhinal – laissant l'hippocampe intact – semblent engendrer des déficits de familiarité, en l'absence de déficit de recollection (Brandt et al., 2016).



Les données de l'imagerie neuroanatomique sont insuffisantes pour établir une cartographie anatomique et fonctionnelle précise des différentes zones impliquées lors de la reconnaissance ; elles ne parviennent pas à rendre compte de toutes les variabilités individuelles dans les déficits mnésiques entre les patients atteints de lésions hippocampiques et/ou parahippocampiques. A titre d'exemple, Holdstock et al. (2002) ont montré un déficit lors de la phase de rappel d'une tâche de reconnaissance Oui/Non chez une patiente présentant une lésion focale au niveau de l'hippocampe, en l'absence de performances déficitaires lors du rappel à une tâche à choix forcés. En outre, d'autres études (Merkow et al., 2015 ; Wais et al., 2006) montrent une implication directe de l'hippocampe dans le processus de familiarité en plus de son rôle dans la recollection, semant le doute sur son implication dans les deux processus de reconnaissance. Finalement, une étude récente (Scalici et al., 2017) montre également l'implication d'autres parties du cerveau, et plus particulièrement des régions préfrontales, dans ces deux processus.

Pris tous ensemble, ces résultats appuient la complexité du phénomène de reconnaissance sur le plan anatomique. Dès lors, la recherche actuelle porte sur l'émergence de modèles intégratifs plus complets, prenant en compte l'intégralité des structures cérébrales, pour expliquer la grande variété de données comportementales qui diffèrent selon les populations (Bastin et al., 2019). Malgré tout, les précédents travaux sur l'hippocampe et les cortex du lobe temporal médian ont permis l'ébauche d'un cadre théorique prometteur, qui semble être compatible avec le modèle DPSD proposé par Yonelinas (1994).

3.4. Données électrophysiologiques

Une dernière constatation concerne les différences de corrélats électrophysiologiques pour les deux processus. En effet, les potentiels évoqués recueillis sur le crâne de nombreux sujets par électroencéphalographie lors d'une tâche de reconnaissance Oui/Non montrent une amplitude plus élevée au niveau fronto-central lorsque les sujets répondent correctement « Non » pour les nouveaux items que lorsqu'ils répondent « Oui » pour les items familiers et similaires (Curran, 2000). Comme ces derniers sont supposés être plus familiers que les nouveaux items correctement rejetés, cette observation est cohérente avec l'activité attendue d'un processus sensible à la familiarité. De plus, l'amplitude au niveau pariétal est supérieure pour les items correctement reconnus par rapport aux items similaires qui ont été reconnus à tort comme étant anciens (Curran, 2000). Cette observation est également cohérente avec l'activité d'un processus lié à la recollection, étant donné que la discrimination entre des mots

extrêmement similaires nécessite le rappel d'informations détaillées. Ces résultats sont par ailleurs appuyés par plusieurs auteurs (Düzel et al., 1997 ; Klimesch et al., 2001), soutenant ainsi l'hypothèse selon laquelle l'activation neuronale des processus de recollection et de familiarité présenterait des différences spatio-temporelles.

Des études électrophysiologiques sur les singes montrent qu'environ 25% des neurones de leur cortex périrhinal répondent plus fortement lorsqu'un nouveau stimulus leur est présenté ; cette activité est plus faible lorsque le même stimulus est présenté une seconde fois (M.W. Brown & Xiang, 1998 ; Xiang & Brown, 1998). Quand un stimulus est présenté plusieurs fois, le signal neuronal de familiarité diminue par ce phénomène de suppression des répétitions (i.e. *repetition suppression*). Ces données renforcent l'importance du cortex périrhinal dans la discrimination entre les anciens et les nouveaux stimuli. Elles sont également à mettre en relation avec celles recueillies dans des populations amnésiques. En effet, chez ces patients, certains neurones réagissant initialement à un stimulus diminuent leur activation de façon permanente tandis que d'autres neurones, approximativement 30%, ne présentent pas cette diminution d'activation (M.W. Brown & Aggleton, 2001 ; M.W. Brown & Xiang, 1998).

4. Évaluation de la familiarité

Ayant ainsi passé en revue les modèles théoriques les plus pertinents pour la modélisation et les arguments en leur faveur, il convient maintenant d'envisager les éléments qui nous permettront de modéliser la familiarité et d'évaluer le potentiel de notre modèle. Ces éléments sont de deux ordres : les tâches de neuropsychologies cliniques et les phénomènes comportementaux.

4.1. Tâches neuropsychologiques

Cette section a pour but de présenter les différents types de tests neuropsychologiques appréciant l'intégrité de la MdR. Dans la pratique clinique, cette dernière est fréquemment évaluée à l'aide de diverses tâches de reconnaissance, qui succèdent à l'apprentissage d'une liste d'images ou de mots. Efficaces dans la pratique, ces méthodes standards ne permettent toutefois pas de quantifier l'implication des processus de recollection et de familiarité à la MdR. Elles n'offrent donc pas une mesure approfondie du SdF (Besson et al., 2012).

4.1.1. Reconnaissance Oui/Non

Le premier paradigme ou « Reconnaissance Oui/Non », tel qu'utilisé dans le test de reconnaissance de visages de la MEM-III (Weschler, 2001), consiste en la présentation d'items

au patient, lequel doit définir si, oui ou non, il les a vus précédemment. A l'entraînement, le patient doit mémoriser une liste d'items présentés visuellement. Dans la phase de reconnaissance, plusieurs items sont présentés un par un au patient. Certains sont connus, d'autres nouveaux. Pour chacun d'entre eux, le patient doit répondre « oui » ou « non » selon qu'il pense avoir ou non déjà étudié cet item lors des phases d'apprentissages précédentes.

4.1.2. Reconnaissance à choix forcés

Le second paradigme ou « Reconnaissance à choix forcés » (RCF), tel qu'utilisé dans le test RL-RI 16 items (Grober & Buschke, 1987), consiste également en la mémorisation d'une liste d'items lors d'une phase d'entraînement. Ensuite, pendant la phase de reconnaissance, le patient se voit successivement présenter des paires d'items, chacune comportant un item précédemment étudié ainsi qu'un nouvel item. Ce dernier a dès lors pour consigne de désigner l'item qu'il a déjà rencontré auparavant, autrement dit l'item familier. En conséquence, les patients amnésiques sont capables de sélectionner l'item correct étant donné qu'ils éprouveront un haut degré de familiarité pour cet item, malgré l'absence d'identification de la trace complète du mot (Grober & Buschke, 1987 ; Holdstock et al., 2002).

4.1.3. Paradigme de mémoire source

Cette approche a pour but d'évaluer dans quelle mesure l'individu se rappelle le contexte de l'encodage d'un item (Wolk et al., 2008). L'individu doit, lors d'une tâche de reconnaissance, désigner la source qui était associée à un item étudié précédemment. Par exemple, lors de l'encodage, l'individu se verra présenter 4 items, disposés aux 4 coins de l'écran. Ce dernier devra, lors de la phase de reconnaissance, déterminer dans quel coin de l'écran était situé l'item, pour autant qu'il ait été correctement reconnu au préalable.

Notons que si cette méthode permet d'évaluer la recollection, elle ne permet pas d'évaluer les processus de familiarité (Besson et al., 2012). Toutefois, un nombre croissant de travaux (Diana et al., 2008 ; Quamme et al., 2007) suggèrent que ce paradigme permet une reconnaissance basée sur la familiarité grâce à l'*unitization* des items, c'est-à-dire qu'un item étudié soit considéré comme un seul et unique élément, ce qui inclut la source associée à ce dernier.

4.2. Phénomènes comportementaux

Trois phénomènes comportementaux liés à la familiarité constitueront nos repères (i.e. *benchmarks*). Ces *benchmarks* seront indicateurs de la qualité de notre modélisation du SdF.

4.2.1. Capacité de stockage illimitée

Alors que la capacité de la mémoire à court terme est généralement estimée à 7 ± 2 items (Miller, 1956), la capacité mnésique en situation de reconnaissance semble, quant à elle, illimitée (Standing, 1973). Dans les expériences de Standing (1973), le SdF a été évalué via la présentation d'un très grand nombre de stimuli, présentés une seule fois pendant 5 secondes chacun. Lors d'un test de reconnaissance, la précision obtenue par les participants, c'est-à-dire le nombre d'images correctement identifiées comme familières, est d'environ 85% lorsque 10000 images naturelles leur sont présentées au préalable. Il n'empêche évidemment que la probabilité d'erreurs croît avec l'augmentation du nombre d'images étudiées.

4.2.2. Effet de récence et de primauté

Lors de ses travaux, Whittlesea (1993) a remarqué que les individus sont capables de distinguer les événements selon qu'ils se sont déroulés dans le passé lointain et le passé récent. L'auteur a donc conduit une expérience afin d'évaluer, à propos d'événements récents, dans quelle mesure les sujets peuvent distinguer ceux ayant eu lieu au début de ceux ayant eu lieu à la fin. Autrement dit, ils ont voulu vérifier si certains items étaient plus facilement reconnus sur base d'un effet de la position de l'item dans l'ordre sériel d'apprentissage. Pour ce faire, les participants ont d'abord visualisé des listes de mots. Ensuite, lors d'une tâche de reconnaissance, ils ont dû indiquer la position chronologique des différentes cibles présentées au sein des listes étudiées précédemment. Whittlesea (1993) a constaté que les premiers et les derniers mots des différentes listes sont plus rapidement placés, suggérant un effet de primauté et de récence. L'effet de primauté semble cependant survenir dans une moindre mesure.

4.2.3. Effet de similarité et fausses reconnaissances

Des études (Hintzman et al., 1992 ; Holdstock et al., 2002 ; Rotello & Heit, 2000) mettent en évidence un effet de la similarité entre un item cible et un leurre sur la MdR. Quand les distracteurs sont similaires aux cibles (une cible serait le mot RAT et un distracteur le mot RATS), les individus se baseront plutôt sur la recollection pour produire leur réponse (Hintzman et al., 1992). Ainsi, une personne avec une atteinte hippocampique, se basant donc uniquement sur un SdF pour réaliser une tâche de reconnaissance, verrait ses performances décliner au fur et à mesure que le chevauchement entre les cibles et les leures augmente. Donc, des distracteurs similaires vont engendrer une forte familiarité (Holdstock et al., 2002), malgré le fait que l'item en question n'a jamais été vu auparavant, ce qui provoque une fausse reconnaissance.

Ayant ainsi posé les bases qui nous permettront de modéliser la familiarité, il convient maintenant de retracer l’historique des principaux modèles informatiques publiés dans la littérature scientifique.

5. Modèles computationnels

Au fil des années 90, plusieurs programmes informatiques se sont essayés à la modélisation de la reconnaissance. Ces modèles, appelés *Global Matching Models*, sont cependant en faveur de l’hypothèse d’un processus unique de force d’un item en mémoire (voir Clark & Gronlund (1996) pour une revue sur le sujet). Dès lors, ils ne permettent ni d’expliquer l’ensemble des données comportementales associées aux TDP, telles que les courbes ROC caractéristiques du processus de recollection, ni de réaliser des prédictions vérifiables concernant les potentiels substrats neuronaux des processus de recollection et de familiarité.

Plus récemment, (McClelland et al., 1995 ; Norman, 2010 ; Norman & O’Reilly, 2003) ont proposé le *Complementary-Learning-Systems model* (CLS), un modèle neuro-computationnel de la MdR qui surmonterait les difficultés rencontrées par les Global Matching Model. En effet, ce modèle serait compatible avec les courbes ROC observées lors de tâches de reconnaissance (Elfman et al., 2008). Il serait également capable de reproduire les distributions du seuil de recollection ainsi que les courbes Gaussiennes du SdF postulées par la DPSD (Yonelinas et al., 2010). Le CLS parvient donc à intégrer les données neuroanatomiques et comportementales concernant la recollection et la familiarité afin d’implémenter leurs contributions à la MdR. Ses principaux fondements ont été élaborés à partir des différences fonctionnelles entre l’hippocampe et les cortex du lobe temporal médian, lesquelles seraient à la base des processus de recollection et de familiarité (McClelland et al., 1995).

Toutefois, ces modèles précurseurs ont pour but la modélisation de la MdR. Or, dans le cadre de ce mémoire, l’objectif n’est pas de reproduire l’ensemble de ce phénomène, en ce compris la recollection et la familiarité, mais bien uniquement le SdF et la reconnaissance basée sur cette familiarité. Nous pouvons néanmoins dégager deux informations essentielles des précédentes tentatives de modélisation. Premièrement, il est possible de reproduire des dissociations comportementales sans utiliser conjointement les deux mécanismes de familiarité et de recollection (Clark & Gronlund, 1996). Deuxièmement, l’indépendance fonctionnelle du cortex périrhinal en situation de reconnaissance par la familiarité nous permet d’envisager la modélisation cognitive de ce processus de façon isolée, en accord avec les TDP (Elfman et al., 2008 ; Norman & O’Reilly, 2003).

Dans la perspective d'un modèle limité au processus de familiarité, deux familles de RNA s'inspirent du fonctionnement des neurones du cortex périrhinal lors de la reconnaissance : les modèles combinés et les modèles spécialisés (Bogacz & Brown, 2003b).

5.1. Modèles combinés

Les modèles combinés supposent que la discrimination sur base de la familiarité ainsi que l'apprentissage des représentations attribuées à un stimulus s'effectuent tous deux au sein d'un seul et même réseau de neurones biologiques. A titre d'exemple, nous retiendrons le modèle CLS proposé par Norman & O'Reilly (2003), le plus souvent cité et qui a l'avantage de montrer des phénomènes comportementaux qui seront explorés dans ce mémoire.

Le modèle CLS (Norman, 2010 ; Norman & O'Reilly, 2003) repose sur l'allégation que l'hippocampe est spécialisé dans la mémorisation rapide d'événements spécifiques tandis que les cortex du lobe temporal médian – ou néocortex – sont, quant à eux, spécialisés dans l'apprentissage lent des régularités statistiques présentes au sein de l'environnement. Le CLS incorpore les aspects biologiques fondamentaux de ces deux systèmes neuronaux en proposant deux modèles computationnels distincts qui permettent la modélisation de la MdR : le modèle hippocampique et le modèle néocortical. La grande différence entre ces deux réseaux réside dans leur aptitude à gérer la séparation des *patterns* qui constituent un stimulus (McClelland et al., 1995 ; Schacter et al., 1998). Dans ce mémoire consacré à la familiarité, seule la partie néocorticale nous intéresse.

La partie néocorticale du CLS (**Figure 5**) part du postulat que le néocortex permet l'association entre des représentations statistiquement similaires à des stimuli précis (Schacter et al., 1998). Concrètement, le chevauchement des caractéristiques structurelles entre une représentation et un stimulus permet au modèle néocortical de reconnaître la structure commune entre eux. Ce modèle ne parvient donc pas à rappeler les détails du stimulus mais bien sa structure globale. Concrètement, lorsqu'un stimulus est présenté à la couche d'entrée du modèle néocortical, ce dernier va créer une représentation interne du stimulus. Au fur et à mesure des expositions répétées, cette représentation interne sera de plus en plus affinée (**Figure 6**). Grâce à un apprentissage Hebbien et à des mécanismes de compétition inhibitrice, l'activité interne va se concentrer sur un plus petit nombre d'unités de la couche cachée, ce qui améliore ainsi ce que les auteurs appellent la netteté (i.e. *sharpness*) de la représentation interne du stimulus. Cette netteté permet au modèle de prendre une décision basée sur la familiarité.

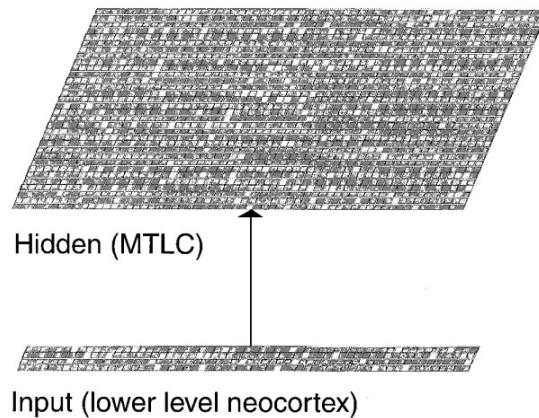


Figure 5 | Structure du modèle néocortical (Norman & O'Reilly, 2003). Le modèle est constitué de deux couches. La couche d'entrée contient 24 lignes de 10 neurones, avec un neurone actif par ligne. La couche cachée contient 1920 neurones reliés par un *feedforward* partiel, dans lequel chaque neurone reçoit des connexions de 25% des entrées. La couche cachée encode les régularités présentes dans la couche d'entrée grâce à un apprentissage Hebbien.

Dans le modèle néocortical, l'apprentissage Hebbien simule trois mécanismes de modifications synaptiques. Premièrement, le mécanisme de *Long-Term Potentiation* (LTP) correspond à l'augmentation des poids entre deux neurones lorsqu'ils sont activés ensemble (Bliss & Collingridge, 1993). Deuxièmement, le mécanisme de *Long-Term Depression* (LTD) correspond à la diminution des poids entre deux neurones lorsqu'un neurone récepteur est activé mais pas le neurone émetteur (Ito, 1989). Troisièmement, le mécanisme de compétition inhibitrice (Grossberg, 1976) est simulé grâce à la règle du *k-Winner-Take-All*, dans laquelle une augmentation de l'activation des neurones gagnants, c'est-à-dire ceux présentant la plus grande activité, va provoquer une diminution de l'activation des perdants.

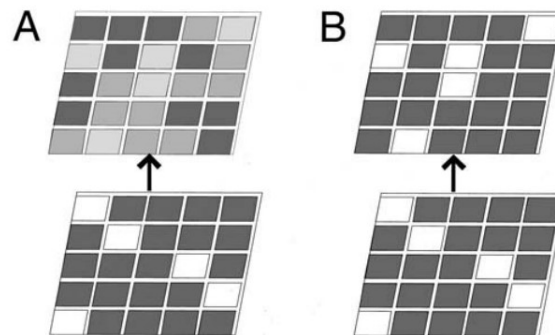


Figure 6 | Représentation interne d'un stimulus, correspondant ici à un *pattern* d'activation aléatoire, (A) avant l'apprentissage et (B) après l'apprentissage (Norman & O'Reilly, 2003).

Afin de simuler le SdF, Norman & O'Reilly (2003) ont émis l'hypothèse qu'un nouveau stimulus activera un plus grand nombre d'unités cachées tandis qu'un stimulus déjà rencontré activera un plus petit nombre d'unités cachées, mais de manière plus importante. Un score de familiarité pour un stimulus sera calculé via l'activité moyenne de ces *top k-units*, c'est-à-dire les k neurones avec le plus haut potentiel membranaire. En fonction d'un certain seuil choisi arbitrairement, le modèle va considérer un stimulus comme étant familier ou non. Le seuil est défini en calculant respectivement la moyenne de l'activité des *top k-units* des cibles et des leurres présentés en phase de test ; il est ensuite placé pile à l'intersection des moyennes (**Figure 7**).

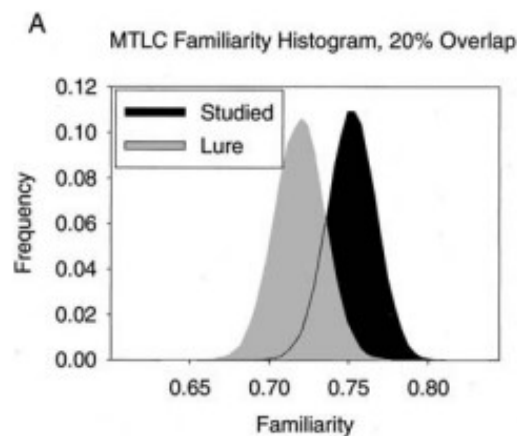


Figure 7 | Distribution des moyennes d'activation des *top k-units* pour les stimuli étudiés lors de l'apprentissage (noir) et les leurres (gris). Le seuil de familiarité est fixé à l'intersection entre les deux courbes (Norman & O'Reilly, 2003).

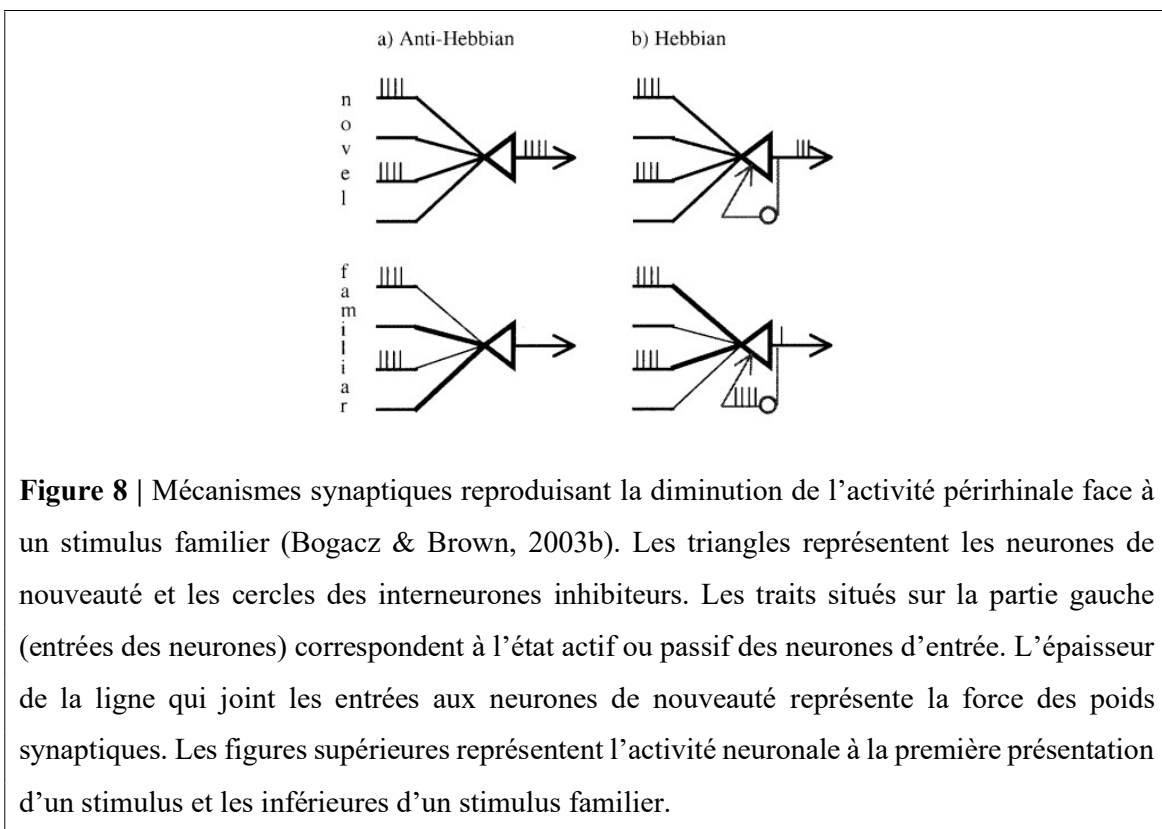
Ici, il est intéressant de constater que la distribution des scores d'activation pour les items cibles et les leurres correspond aux deux courbes Gaussiennes observées dans la DPSD (voir **Figure 2**). Ces deux courbes se chevauchent fortement, certains leurres étant très similaires aux cibles d'un point de vue morphologique. Ainsi, ces leurres vont déclencher une forte familiarité augmentant la probabilité d'une erreur de reconnaissance.

Le modèle néocortical original a été conçu à partir de synapses complexes (i.e. *Spiking Neurons* ou neurones continus) ; chaque neurone d'entrée possède sa propre valeur d'activation – ou potentiel membranaire – calculée grâce à une fonction d'activation spécifique. Néanmoins, Bogacz & Brown (2003b) propose une version simplifiée, à l'aide de RNA classiques, dans laquelle l'activation des neurones est binaire : les neurones actifs prennent la valeur de 1 et les neurones inactifs celle de 0.

5.2. Modèles spécialisés

Les modèles spécialisés partent du principe qu'une petite proportion des neurones périrhinaux constituent un réseau spécialisé dans la seule reconnaissance sur base de la familiarité (Bogacz et al., 2001b ; Bogacz & Brown, 2003a). Ces neurones de nouveauté (i.e. *novelty neurons*) constituent approximativement 10% de l'ensemble des neurones périrhinaux et ont la particularité de s'activer de façon importante lors de la présentation d'un nouveau stimulus. Si ce stimulus est visionné une seconde fois, les neurones de nouveauté répondront brièvement et plus faiblement (M.W. Brown & Xiang, 1998 ; Xiang & Brown, 1998).

Dans ces modèles, chaque stimulus est représenté par un *pattern* spécifique d'activités binaire, c'est-à-dire actif ou inactif, constituant l'entrée du réseau. L'activité de ces neurones d'entrée correspond donc aux caractéristiques des stimuli (Bogacz & Brown, 2003b). Ces réseaux composés de deux couches sont donc utilisés pour apprendre et mémoriser des *patterns* d'activités. En outre, ils parviennent à mimer la diminution de l'activité des neurones du cortex périrhinal lors de la présentation d'un stimulus familier (Bogacz & Brown, 2003a, 2003b). Néanmoins, chaque réseau spécialisé diffère en termes de plasticité synaptique, c'est-à-dire dans les modifications des poids entre les couches (**Figure 8**). Dans cette section, deux modèles de plasticité synaptique seront explorés : le modèle anti-Hebbien et le modèle Hebbien.



Il a été démontré par des méthodes analytiques et des simulations informatiques (Bogacz & Brown, 2002) que la capacité de la mémoire pour la reconnaissance de la familiarité fournie par certains de ces modèles est de l'ordre n^2 , où n est le nombre d'unités dans le réseau neuronal. Ces résultats peuvent être considérés comme une explication mathématique de l'immense capacité de mémoire lors de la reconnaissance par la familiarité (Standing, 1973). Les modèles spécialisés montrent ainsi des capacités de mémoire largement supérieures à celles des modèles combinés (Bogacz & Brown, 2003b ; Norman & O'Reilly, 2003).

5.2.1. Modèle anti-Hebbien

La **Figure 8A** représente un neurone du modèle anti-Hebbien (Bogacz & Brown, 2003a), basé sur des diminutions plutôt que sur des augmentations de la force entre les synapses. Après la présentation d'un nouveau stimulus, les forces des synapses connectées avec les neurones d'entrée actifs sont diminuées, reproduisant le mécanisme de LTD homo-synaptique. Cette modification synaptique diminue la somme des poids synaptiques du neurone de nouveauté. Par conséquent, pour maintenir l'excitabilité globale du neurone, les poids synaptiques des connexions des neurones d'entrée inactifs doivent être augmentés. Lorsque le même stimulus est présenté à nouveau, le potentiel membranaire du neurone de nouveauté sera plus faible car les poids synaptiques des entrées qui étaient activées pour ce stimulus ont été réduits ; le neurone de nouveauté sera généralement moins actif, voire inactif. Ainsi, le neurone répond plus fortement aux nouveaux stimuli qu'aux stimuli familiers.

En résumé, les connexions des entrées actives aux neurones de nouveauté sont diminuées comme si cela était dû à la dépression à long terme (LTD) ; ainsi, sans avoir besoin d'inhibition, les neurones de nouveauté répondent moins fortement à un stimulus s'il a déjà été vu par le réseau. Ceci correspond aux preuves expérimentales démontrant une réduction de l'activité des neurones du cortex périrhinal face à un stimulus familier (M.W. Brown & Xiang, 1998 ; Xiang & Brown, 1998).

5.2.2. Modèle Hebbien

La **Figure 8B** montre le modèle Hebbien (Bogacz et al., 2001b), basé sur la plasticité synaptique du même nom. Ici, les connexions entre les entrées actives et les neurones de nouveauté sont augmentées via un processus semblable à la potentialisation à long terme (LTP). De la sorte, un neurone de nouveauté est plus susceptible de se déclencher, dans la période qui suit directement l'apparition du stimulus, pour un stimulus familier que pour un nouveau stimulus. Cependant, les neurones de nouveauté sont supposés répondre moins fort pour un

stimulus familier (M.W. Brown & Xiang, 1998 ; Xiang & Brown, 1998). Pour pallier ce problème, ces derniers reçoivent une entrée inhibitrice qui supprime leurs réponses. Étant donné que cette inhibition est entraînée par les réponses initiales des neurones de nouveauté eux-mêmes, ces réponses initiales sont dès lors supprimées pour les items familiers et non pour les nouveaux. Ainsi, dans la période qui suit les réponses initiales, la présentation d'un stimulus engendre une réponse neuronale plus forte quand le stimulus est nouveau que quand il est familier. Les auteurs (Bogacz et al., 2001b ; Cowell, 2012) suggèrent que, dans les neurones de nouveauté observés expérimentalement, l'activité dans la phase post-stimulus initiale et brève est masquée par l'activité dans la phase inhibitrice ultérieure, plus longue. De la sorte, le déclenchement observé des neurones de nouveauté dans le cortex périrhinal ressemble à celles des neurones de nouveauté dans le modèle.

En résumé, après la présentation d'un nouveau stimulus, les poids synaptiques des neurones d'entrée actifs sont augmentés, mimant ainsi le mécanisme de LTP homo-synaptique ; les poids des unités inactives sont également diminués comme s'ils étaient modifiés par les mécanismes de LTD hétéro-synaptique ou homo-synaptique. Ces changements produisent une réponse des neurones de nouveauté initialement plus élevée pour les stimuli familiers que pour les nouveaux. Pour correspondre à la réduction d'activité observée dans les neurones biologiques du cortex périrhinal (M.W. Brown & Xiang, 1998 ; Xiang & Brown, 1998), les neurones de nouveauté du modèle Hebbien vont projeter sur un interneurone inhibiteur. En retour, celui-ci va augmenter le niveau d'inhibition pour les stimuli familiers. Cette inhibition accrue va réguler l'activité des neurones de nouveauté de sorte qu'ils produisent une réponse plus faible pour ces stimuli familiers. A l'inverse, les nouveaux stimuli présenteront un faible niveau d'inhibition et auront donc une réponse plus forte de la part des neurones de nouveauté.

5.3. Limitation de l'activité des neurones de nouveauté

Dans un RNA plus complet modélisant la familiarité, une couche complète de neurones de nouveauté reçoit des projections d'une première couche de neurones d'entrée (Androulidakis et al., 2008). Autrement dit, il s'agit d'un réseau « deux-couches » à propagation avant (i.e. *feedforward*, **Figure 9**). Si chaque neurone de nouveauté prend sa propre décision sur la familiarité du stimulus, c'est la réponse globale du réseau qui permet la décision finale de familiarité, c'est-à-dire l'ensemble de l'activité des neurones de nouveauté (Bogacz & Brown, 2003b). Le postulat est que les stimuli familiers provoqueront une activité moyenne plus faible que les nouveaux (Androulidakis et al., 2008 ; Bogacz & Brown, 2003a).

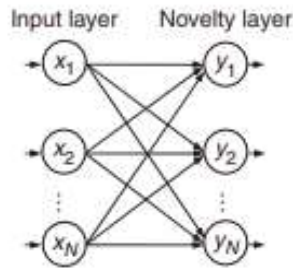


Figure 9 | Architecture d'un réseau « deux couches » *feedforward* (Androulidakis et al., 2008). Les cercles représentent les neurones et les flèches les connexions. Le réseau est entièrement connecté, avec une couche de neurones de nouveauté (y_i) qui reçoivent des projections d'une couche de neurones d'entrée (x_i).

Pour maximiser la capacité de stockage d'informations du réseau, il est nécessaire de s'assurer que chaque neurone de nouveauté reste un évaluateur indépendant de la familiarité d'un stimulus (Bogacz et al., 2001b). En effet, si tous les neurones de nouveauté étaient actifs après la présentation de chaque nouveau stimulus, alors tous les poids seraient modifiés de la même manière et seraient ainsi fortement corrélés. Les neurones de nouveauté seraient donc tous actifs (ou inactifs) en même temps et l'ensemble du réseau fonctionnerait comme un seul et unique neurone. Pour éviter ce problème, le nombre de neurones de nouveauté actifs pour un stimulus donné doit être limité.

5.3.1. Compétition inhibitrice

Pour limiter le nombre de neurones de nouveauté actifs, le modèle anti-Hebbien recourt à des mécanismes de compétition inhibitrice (Grossberg, 1976). A cet effet, seule la fraction de neurones ayant les potentiels membranaires les plus élevés est sélectionnée pour être active ; l'activité du reste des neurones est supprimée par inhibition (Bogacz & Brown, 2003b). Seuls les neurones gagnants, c'est-à-dire les plus actifs, participeront à la modification des poids. Les neurones perdants n'y participeront pas. Cette méthode de limitation du nombre de neurones actifs est également utilisée dans les modèles combinés (Norman & O'Reilly, 2003).

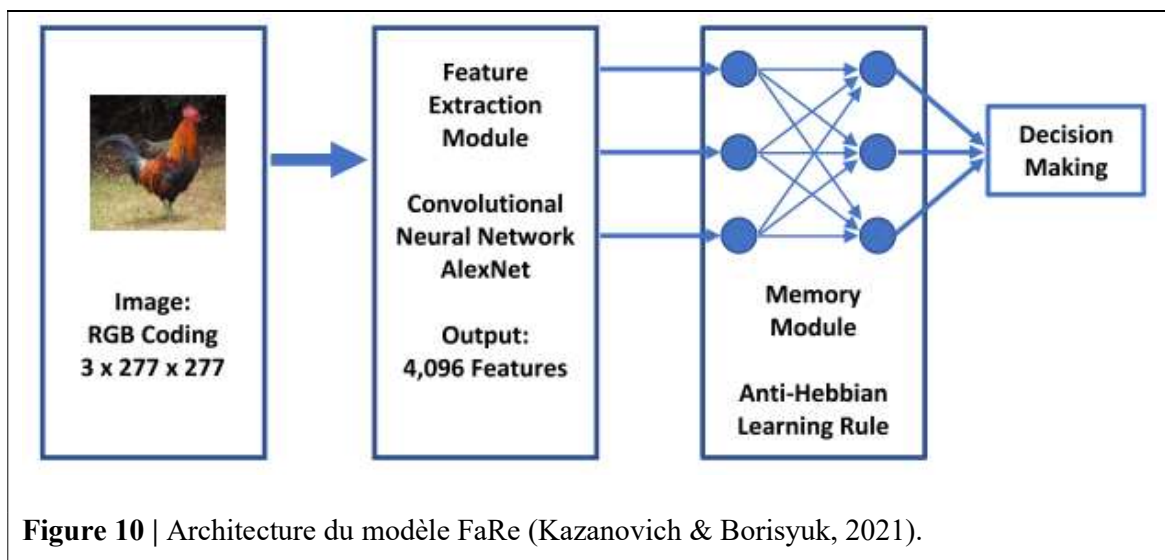
5.3.2. Fortes connexions

Il existe d'autres façons de limiter le nombre de neurones de nouveauté en activité lors de la présentation d'un stimulus. Dans le modèle Hebbien, ce problème est résolu en fournissant au réseau des connexions spécifiques présentant des poids synaptiques élevés entre certaines entrées du réseau et un sous-ensemble de neurones de nouveauté (Bogacz & Brown, 2003b).

Ces poids synaptiques ne seront, par ailleurs, pas modifiables lors de la mise à jour du réseau. Ainsi, les neurones de nouveauté sélectionnés comme étant actifs sont ceux pour lesquels il existe une forte connexion (non-modifiable) avec une entrée correspondante, elle-même à l'état actif. Les neurones de nouveauté ne présentant aucune forte connexion avec un neurone d'entrée actif seront dès lors au repos.

5.4. Modèle FaRe

La plupart des modèles qui viennent d'être exposés se nourrissent d'entrées binaires (Androulidakis et al., 2008 ; Bogacz & Brown, 2003a), c'est-à-dire des suites de 0 et de 1, ou encore de *patterns* d'activités aléatoires en guise de stimuli à mémoriser (voir **Figure 6**). Dès lors, il existe toujours un fossé entre le fonctionnement de ces modèles artificiels et celui du cerveau humain. Il est effectivement plus difficile de réaliser un jugement de familiarité avec des formes abstraites (Bellhouse-King & Standing, 2007). Il serait donc légitime de se questionner quant à la survenue d'une expérience de familiarité sur base abstraite dans la vie réelle. Bien que ces mécanismes soient encore mal connus, il semble de même peu probable qu'une situation de la vie quotidienne soit traitée par notre cerveau sous la forme d'une suite binaire (Eichenbaum et al., 2007). Récemment, deux équipes de chercheurs (Ji-An et al., 2022 ; Kazanovich & Borisyuk, 2021) ont essayé de combler ces lacunes en fournissant les caractéristiques d'images naturelles en entrées du réseau. Le fonctionnement de ces deux modèles étant assez similaires, nous retiendrons spécialement le modèle FaRe (**Figure 10**), implémenté par Kazanovich & Borisyuk (2021), qui a servi de base à la conception de notre modèle.



Le *Familiarity Recognition model* (FaRe) est prometteur car ses résultats correspondent à certains phénomènes comportementaux liés au SdF. Effectivement, il parvient à reproduire la capacité de stockage illimitée pour la reconnaissance basée sur la familiarité avec des images naturelles (Standing, 1973). Il est également moins performant dans le cas d'images abstraites ou bruitées, ce qui correspond aux données expérimentales (Bellhouse-King & Standing, 2007).

Le modèle FaRe est un RNA qui combine un réseau profond avec un apprentissage anti-Hebbien afin de modéliser le SdF sur des images naturelles (Kazanovich & Borisyuk, 2021). Comme illustré sur la **Figure 10**, l'architecture du modèle est composée de deux modules. La phase d'apprentissage (i.e. *training* ou entraînement) se déroule donc en deux étapes.

La première étape reproduit le traitement de l'information par les régions associatives visuelles jusqu'au cortex périrhinal, fournissant des *patterns* d'entrées qui seront traités ultérieurement lors de la deuxième étape. Dans cette première étape, les caractéristiques d'une image sont extraites par un réseau profond convolutif (RPC) pré-entraîné. Les détails de ce module d'extraction sont décrits dans l'**Annexe 1**. Les auteurs (Kazanovich & Borisyuk, 2021) ont conclu qu'une décision correcte de familiarité dépend entre autres de l'entraînement préalable du réseau profond utilisé. Selon eux, cela rend compte du développement précoce du système visuel lors de l'enfance, qui se forme de façon à permettre l'encodage efficace des images de la vie de tous les jours.

Lors de la deuxième étape, un réseau neuronal *feedforward* composé de deux couches est utilisé pour la familiarité. Ce réseau s'inspire directement du modèle anti-Hebbien décrit ci-dessus afin de modéliser le fonctionnement du cortex périrhinal. (Androulidakis et al., 2008 ; Bogacz & Brown, 2003b, 2003a). Rappelons que la règle d'apprentissage anti-Hebbien permet une diminution de l'activité des neurones de nouveauté de la couche de sortie pour un stimulus devenu dès lors familier. Les détails mathématiques de ce module de mémoire sont décrits dans l'**Annexe 2**.

Notons cependant que, contrairement aux modèles précédents, aucun mécanisme de compétition inhibitrice ne semble avoir été implémenté dans le modèle FaRe afin de limiter le nombre de neurones de nouveauté actifs. L'article sous-entend que de tels mécanismes entrent en jeu dans le modèle FaRe (Kazanovich & Borisyuk, 2021, p. 630) mais ne donne aucune indication mathématique sur la manière dont ils auraient été implémentés. Finalement, nous avons cru comprendre qu'ils n'ont pas été inclus lors de la modélisation (R. Borisyuk, communication personnelle).

Focus. Traitement des informations visuelles jusqu'au cortex périrhinal.

Le cortex périrhinal reçoit des afférences de plusieurs régions sensorielles. Les informations reçues, notamment depuis les aires visuelles, sont largement traitées avant d'arriver jusque-là (Eichenbaum et al., 2007). Il n'existe à ce jour aucune description précise des caractéristiques qui constituent un stimulus visuel à l'entrée de cette structure corticale. Néanmoins, il semble concevable qu'elles soient formées dans les régions cérébrales impliquées dans le développement précoce du système visuel chez l'homme (Kazanovich & Borisyuk, 2021).

Le système visuel transforme la lumière entrante en représentations significatives qui vont soutenir sa perception. Cette transformation se produirait au travers d'une cascade de processus hiérarchiques, ces derniers mis en œuvre dans un ensemble de régions du cerveau le long des flux visuels ventral et dorsal (Humphreys & Riddoch, 2006). Chacune de ces régions semble remplir des sous-fonctions distinctes afin de permettre la perception visuelle. Ainsi, la majeure partie des projections néocorticales vers le cortex périrhinal proviendraient de régions associatives. Ces afférences traitent l'information sensorielle unimodale relative à la qualité de l'objet (« What » ou « qu'est-ce que c'est ») en remontant le long du flux visuel ventral (Eichenbaum et al., 2007).

La phase de test (i.e. testing) correspond à une tâche de RCF durant laquelle des paires d'images sont montrées successivement au modèle. Chaque paire d'images est composée d'une image apprise lors de l'entraînement ainsi que d'une nouvelle image. En accord avec les données anatomiques et neurophysiologiques (M.W. Brown & Xiang, 1998 ; Eichenbaum et al., 2007 ; Xiang & Brown, 1998), une image familière devrait présenter un potentiel membranaire moins élevé qu'une image qui n'a jamais été apprise par le modèle. La fonction de décision est basée sur ce postulat, c'est-à-dire qu'une image familière présentera en moyenne une activité moindre sur la couche de sortie qu'une nouvelle image.

6. Objectifs du mémoire

Dans ce mémoire, nous avons programmé un modèle connexionniste du SdF, c'est-à-dire un modèle informatique qui utilise des RNA afin de reproduire la discrimination sur base de la familiarité lors de la réalisation d'une tâche de reconnaissance. Plus précisément, nous avons implémenté une variante du modèle FaRe proposé par Kazanovich & Borisyuk (2021). En effet, ce modèle est le seul, à notre connaissance, qui utilise de simples neurones artificiels, comme ceux décrits par McCulloch & Pitts (1943), pour modéliser la familiarité vis-à-vis d'images naturelles et non d'entrées binaires (Androulidakis et al., 2008).

Dans notre variante, nous utilisons un modèle d'extraction de caractéristiques plus récent que le réseau AlexNet (Krizhevsky et al., 2012), utilisé dans le modèle FaRe. En effet, il est probable que l'entraînement préalable du réseau convolutif utilisé joue un rôle déterminant dans la réussite des tâches de reconnaissance (Kazanovich & Borisyuk, 2021). L'implémentation d'un RPC plus performant pourrait dès lors améliorer l'ajustement entre les performances du modèle et les données de la littérature. C'est donc une version pré-entraînée du réseau convolutif ResNet50 (He et al., 2015, 2016) que nous avons utilisée dans notre modèle.

Ensuite, nous utilisons également un module de mémoire Hebbien comme alternative au module anti-Hebbien du modèle FaRe. Nous avons ainsi adapté l'apprentissage Hebbien décrit par Bogacz & Brown (2003b) pour des entrées naturelles et non binaires. Cette nouvelle règle d'apprentissage, en plus d'être biologiquement plus plausible que la règle anti-Hebbienne, parvient à reproduire avec succès l'activité périrhinale des stimuli devenus familiers présentant des entrées binaires (Bogacz et al., 2001b ; Bogacz & Brown, 2003b).

En plus de reproduire les résultats obtenus par les chercheurs (Kazanovich & Borisyuk, 2021 ; Standing, 1973), nous avons simulé une tâche de reconnaissance dans laquelle les stimuli cibles sont similaires aux leurres sur les plans perceptif et conceptuel (Holdstock et al., 2002). Ainsi, nous avons fait la prédiction que lorsque la cible et le leurre font partie de la même catégorie sémantique, l'importance du degré similarité entre les items fait chuter les performances du modèle et augmenter la probabilité d'erreurs (Norman & O'Reilly, 2003).

Nous avons également exploré un phénomène observé par Ji-An et al. (2022) dans leur modèle complexe sur la détection de visage familier. Ainsi, après l'apprentissage d'une série de visages, nous avons réalisé une tâche de reconnaissance dans laquelle les cibles ont été présentées selon d'autres perspectives. Nous pouvons ainsi vérifier si un modèle réalisé à partir

de simples RNA est capable de généraliser la familiarisation d'un visage sur différentes tranches d'âge ou encore avec une pilosité différente. En effet, dans la vie de tous les jours, il semble peu probable qu'une personne ne puisse reconnaître quelqu'un comme étant familier en raison de caractéristiques physiques différentes que lors de la précédente rencontre.

En parallèle, nous avons joué avec les différents paramètres du modèle afin d'en évaluer l'impact sur les performances lors de tâches de reconnaissance. Plus concrètement, ont été explorés : le nombre de neurones de nouveauté en sortie, la constante d'apprentissage (i.e. *learning rate*) ou encore la tâche de reconnaissance utilisée.

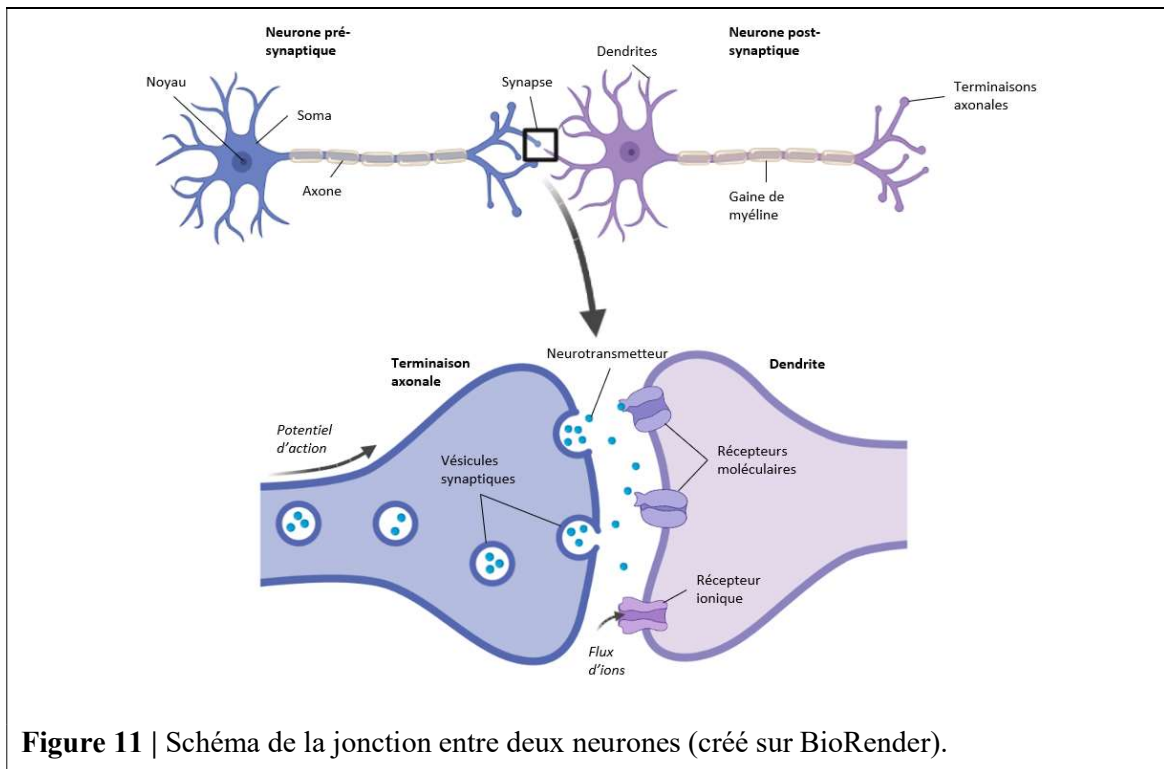
Nous avons également comparé les performances de notre modèle par rapport à celles du modèle FaRe original, proposé par Kazanovich et Borisyuk (2021), lors des tâches de reconnaissance précédemment réalisées. Conformément aux observations de Bogacz & Brown (2003b) quand ils ont eux-mêmes comparé les performances de ces différents modèles, nous nous attendions à une capacité de mémoire plus faible dans notre nouveau module de mémoire ; nous avons néanmoins fait l'hypothèse que les performances de notre modèle rendraient mieux compte d'un effet de la similarité des stimuli. L'implémentation de l'architecture FaRe a été réalisée sur Python. Les détails de cette modélisation sont disponibles dans les **Annexes 1 et 2**.

PARTIE II : MODÉLISATION

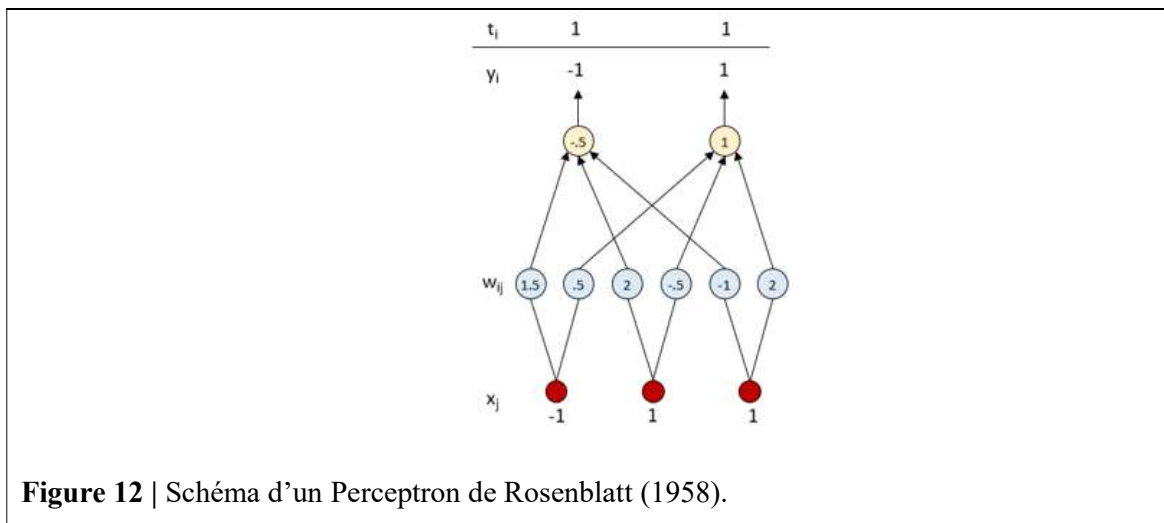
Dans cette seconde partie, nous présenterons le modèle connexionniste que nous avons conçu afin de reproduire artificiellement le SdF. Son fonctionnement est fortement similaire à celui de son prédécesseur, le modèle FaRe (Kazanovich & Borisjuk, 2021), présenté dans la section précédente. Ensuite, nous présenterons les différentes simulations réalisées ainsi que leurs détails méthodologiques. Les résultats de ces simulations seront également exposés. Cependant, avant d’aller plus loin dans l’exploration de notre modèle, il est nécessaire de bien comprendre le fonctionnement d’un neurone artificiel.

1. Réseaux de neurones artificiels

Un neurone biologique est composé d’un soma, qui se prolonge par un axone, ce dernier se divisant lui-même en plusieurs terminaisons axonales (**Figure 11**). Les terminaisons de ce neurone dit pré-synaptique vont venir toucher les dendrites (ou le soma) d’un autre neurone appelé cette fois « neurone post-synaptique ». Ce point de contact entre deux neurones s’appelle une synapse. Le neurone pré-synaptique va modifier l’état du neurone post-synaptique : quand le premier émet son potentiel d’action, c’est-à-dire se décharge électriquement, celui-ci parcourt l’axone jusqu’aux terminaisons, où des neurotransmetteurs vont se libérer dans la synapse. Ceux-ci se fixent sur des récepteurs post-synaptiques, ce qui va provoquer l’ouverture de canaux ioniques. Le message électrique va dès lors pouvoir entrer dans la cellule post-synaptique et modifier le potentiel membranaire initial du neurone post-synaptique.



Les réseaux de neurones artificiels ont été pensés à l'image du fonctionnement de ces neurones biologiques. McCulloch & Pitts (1943) sont les premiers à avoir tenté de reproduire artificiellement la mécanique des neurones telle que décrite ci-dessus. Plus tard, Rosenblatt (1958) améliore leur modèle et imagine le perceptron (**Figure 12**). Il s'agit d'un dispositif avec toute une série d'entrées (x_i) connectées à des nœuds. Les liens entre les entrées et un nœud sont appelés poids (w_{ij} , i.e. *connection strenghts*). La valeur d'un nœud correspond simplement à la somme pondérée des poids multipliés par l'entrée correspondante. Ensuite, comme exposé par Defays et al. (1997), « ce nœud répondra positivement ou négativement suivant que cette somme est supérieure ou inférieure à un seuil donné », c'est-à-dire qu'il prendra respectivement la valeur de 1 ou -1. C'est le principe de base du concept de *forward propagation* ou propagation vers l'avant. Ces valeurs de sortie (y_i) sont finalement comparées aux valeurs désirées (t_i) en sortie de perceptron.



Pour qu'il y ait un apprentissage, les poids devront subir une modification progressive pour que les valeurs de sortie du perceptron correspondent au plus près aux valeurs désirées. En 1949, Donald Hebb postulait que « quand un axone d'une cellule A excite une cellule B et que, de manière répétée et persistante, il prend part à son déclenchement, un processus de croissance ou un changement métabolique survient dans l'une ou les deux cellules de telle façon que l'efficacité de A, en tant que cellule provoquant la décharge de B, est augmentée ». En d'autres termes, lorsque deux neurones sont excités conjointement, le lien qui les unit est renforcé. Ce postulat est à l'origine de l'apprentissage Hebbien et a été réutilisé par Rosenblatt (1958) pour entraîner son perceptron, en ajustant les poids w_{ij} . La fonction d'apprentissage permet de modifier le poids des liens entre les unités d'entrée et de sortie comme suit :

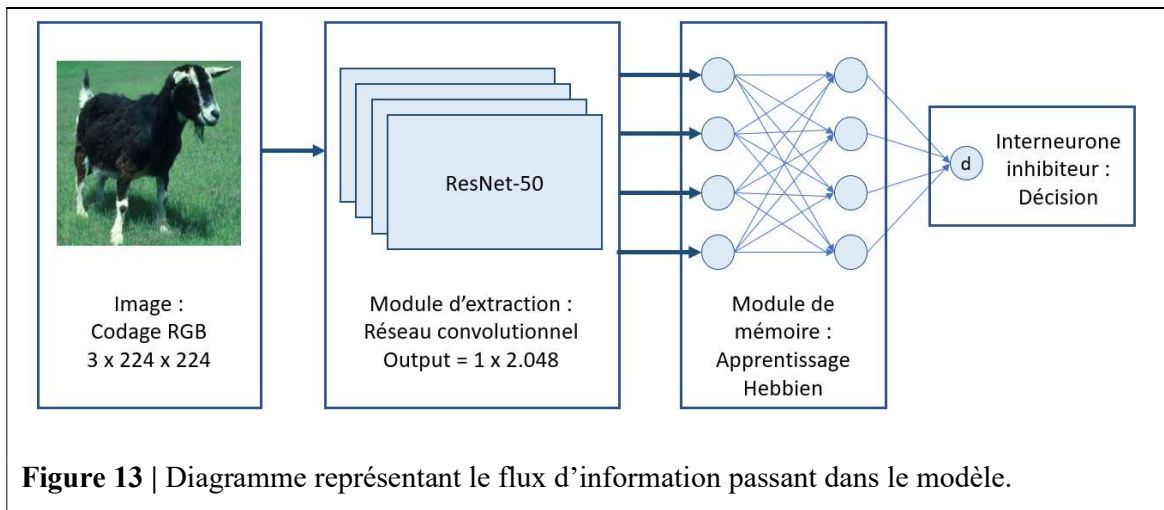
$$w_{ij(nouveau)} = w_{ij(ancien)} + \eta y_i x_j$$

où η correspond à la constante d'apprentissage qui va venir conditionner la vitesse à laquelle les poids seront modifiés.

2. Architecture du modèle

Notre modèle présente une architecture combinant des couches de convolution avec apprentissage Hebbien (**Figure 13**). Comme nous l'avons déjà exposé, son fonctionnement est similaire à celui du modèle FaRe, dont l'efficacité du mode de fonctionnement en deux étapes pour effectuer un jugement de familiarité sur des images naturelles a déjà été démontrée (Ji-An et al., 2022 ; Kazanovich & Borisyuk, 2021). Il s'agit d'un RNA qui reprend les conceptions théoriques du fonctionnement du cortex périrhinal afin de simuler un jugement de reconnaissance sur base de la familiarité d'un stimulus. L'implémentation du modèle a été effectuée sur Python 3.9.11 ; le code est disponible en open access à l'adresse suivante :

<https://github.com/JRead98/master.git>.



Le processus de détection de familiarité s'effectue donc en deux temps. Lors de la première phase, une image est traitée par le RPC pré-entraîné ResNet50 (He et al., 2015, 2016), afin d'en extraire les caractéristiques (i.e. *features*). Cette première étape, appelée module d'extraction, a pour objectif de mimer le traitement de l'information depuis les aires visuelles associatives du cerveau jusqu'au cortex périrhinal. Elle permet la production des entrées du réseau, constituées de nombres réels. Le vecteur de sortie du modèle d'extraction correspond ainsi aux caractéristiques d'une image donnée, qui sera traité ultérieurement par les neurones du module de mémoire.

Focus. Les RPC pour mimer le fonctionnement des aires visuelles

Les RPC, autrement appelés ConvNet, consistent en la succession de couches de convolution, fonctions de transfert ReLU et couches d'agrégation (i.e. *pooling*), programmées pour reproduire la détection des traits constitutifs d'un stimulus par les cellules du cortex visuel (le Cun, 2019). Dans ce type de réseau, un algorithme de rétropropagation du gradient permet l'ajustement des poids pour que les neurones des différentes couches parviennent à détecter des motifs particuliers (barres verticales ou horizontales, couleurs, ...). Par exemple, dans la première couche de convolution, qui correspond à la couche visuelle V1, les cellules simples vont détecter une ligne d'orientation particulière, comme une barre verticale, quel que soit l'endroit où ce trait figure sur l'image.

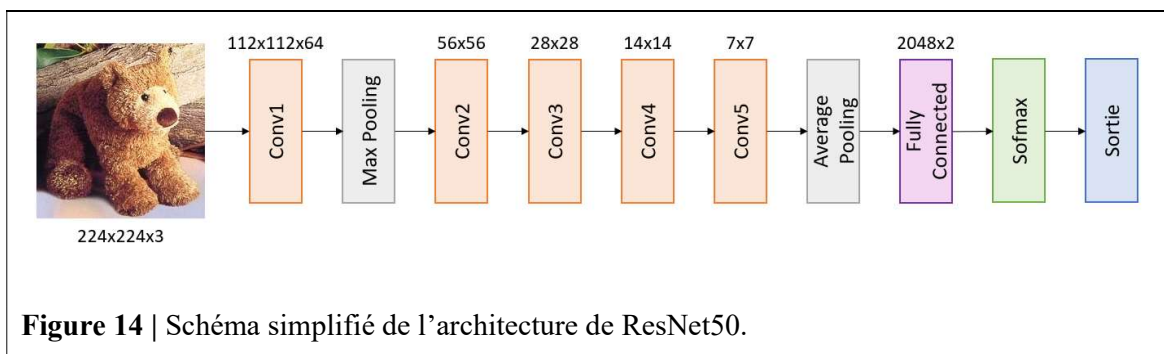
La convolution est une opération mathématique qui va produire une carte de caractéristiques (i.e. *feature map*), via la somme pondérée de tous les neurones, signalant par exemple la présence ou l'absence de barres verticales. Cette *feature map* est ensuite passée dans une fonction d'activation ReLU. Celle-ci est choisie pour sa simplicité mathématique ($y = x$ si x est positif, sinon 0) et sa capacité à éviter la saturation du gradient (Krizhevsky et al., 2012). Le résultat est ensuite passé dans une couche de *pooling*. Il s'agit d'une opération analogue à celle de l'agrégation qui est réalisée par les cellules complexes du cortex visuel. Elle permet de produire une représentation invariante du trait en cas de petits déplacements de ce trait dans l'image. Dans une couche de convolution, un empilement de *feature maps* permet de détecter la présence de plusieurs motifs particuliers, tels que les oreilles pointues d'un chat.

La seconde phase utilise les vecteurs de caractéristiques obtenus par le module d'extraction afin de modéliser la familiarité de l'image présentée au modèle. Pour ce faire, les entrées – c'est-à-dire les sorties du RPC - passe dans un simple réseau de neurones composé de deux couches, relativement similaire à ceux utilisés dans les travaux précédents (Androulidakis et al., 2008 ; Bogacz & Brown, 2003a). Ensuite, les *features* de l'images sont mémorisées dans ce module de mémoire grâce à un apprentissage Hebbien (Bogacz et al., 2001b). Rappelons que ce type d'apprentissage semble biologiquement plus plausible que son homologue anti-Hebbien (Bogacz & Brown, 2003b). Initialement pensé avec des réseaux de neurones continus, cet apprentissage Hebbien a été adapté à des réseaux de neurones classiques tels que décrits précédemment. Cette simplification lui permet également de fonctionner avec des entrées naturelles. En effet, Kazanovich et Borisjuk (2021) ont déjà reproduire des données comportementales sans avoir eu recours à ce type de neurones complexes.

2.1. Module d'extraction des caractéristiques

Afin de reproduire le traitement de l'information visuelle jusqu'au cortex périrhinal, nous avons décidé d'utiliser le vecteur de caractéristiques tel qu'il apparaît à l'une des couches supérieures d'un RPC pré-entraîné. La couche choisie correspond à l'encastrement des nombreuses couches de convolution successives qui constituent le RPC. En effet, ce type d'architecture a été conçu à l'origine pour reproduire le traitement de l'information dans les aires visuelles qui se succèdent, à l'image des différentes couches de convolution d'un RPC (le Cun, 2019).

Dans notre variante du modèle FaRe, le module d'extraction est un RPC du nom de ResNet50 (He et al., 2016). Plus précisément, il s'agit de ResNet50 v1.5 (He et al. (2015), qui a été préalablement entraîné sur PyTorch avec 1.2 millions de photographies haute résolution d'images naturelles provenant d'ImageNet. Il a été initialisé comme décrit dans l'article de He et al. (2015). A l'origine, ResNet50 permet la classification d'images dans 1000 catégories différentes avec un taux d'erreurs inférieur à 4,94%. Un diagramme simplifié de l'architecture de ResNet50 est représenté sur la **Figure 14**.

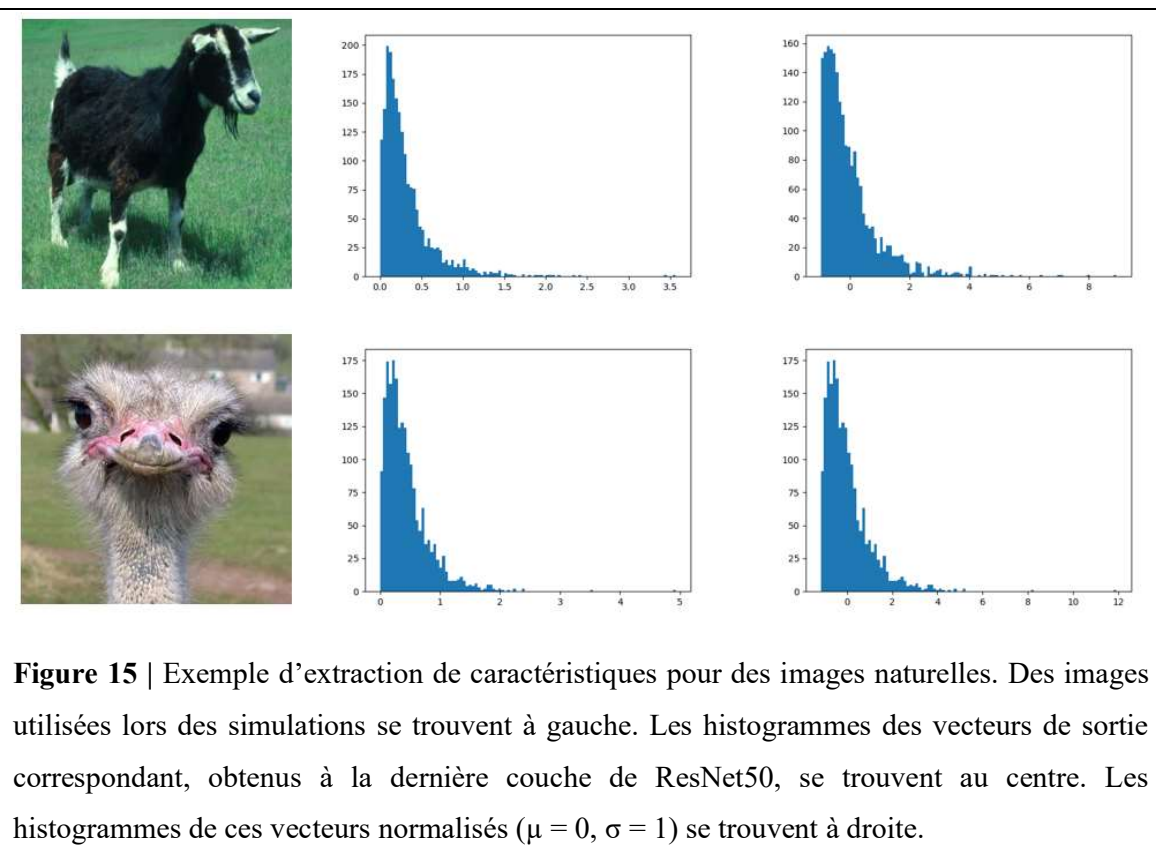


Ce RPC est constitué de 48 couches de convolution et 2 couches de *pooling*, pour un total de 50 couches permettant d'identifier une image et d'en définir ses caractéristiques selon différents degrés de complexité. L'avant-dernière couche du modèle, située à la suite de la seconde couche de *pooling*, est une couche entièrement connectée (i.e. *fully-connected*) de taille 2048. Nous utilisons cette couche pour représenter les caractéristiques d'une image.

Dans l'architecture complète de ResNet50, la couche *fully-connected* projette sur une couche de *softmax*, qui permet au réseau de classer l'image qui lui est présentée. Nous n'utilisons pas cette *softmax* dans notre implémentation. En plus d'avoir une précision supérieure aux autres RPC, ResNet50 permet de résoudre le problème de dégradation, fréquent lorsque la profondeur de ce type de réseau augmente, et ce, grâce à des connexions raccourcies

(i.e. *shortcut connections*). Ces dernières ne sont pas représentées sur le graphique ci-dessus.

Avant de passer dans le module d'extraction, la représentation RGB de chaque image a été normalisée à la taille 3x224x224, 3 étant le nombre de couches correspondant aux couleurs RGB et 224x224 correspondant à la taille des images. Pour une image donnée, nous avons ensuite récupéré le vecteur obtenu à l'avant-dernière couche *fully-connected* de ResNet50, que nous considérons comme étant les *features* de l'image. Ce vecteur est utilisé pour l'apprentissage de l'image dans le module de mémoire. Après être passé dans le réseau convolutif, le vecteur de taille 2048 pour une image donnée, correspondant à ses caractéristiques, est récolté avant d'être normalisé, c'est-à-dire que la distribution des valeurs du vecteur présente une moyenne de 0 et un écart-type de 1. Nous avons utilisé ce vecteur de nombres réels comme entrées pour le module de mémoire. La **Figure 15** montre les histogrammes des valeurs du vecteur de sortie obtenus après cette première étape de notre modèle.



2.2. Module de mémoire Hebbien

Le module de mémoire Hebbien comporte deux couches reliées par un réseau *feedforward* (voir **Figure 9**). L'entraînement du module s'effectue via une règle d'apprentissage Hebbienne (Hebb, 1949) adaptée par nos soins aux nombres réels obtenus à la sortie du RPC. Les détails d'un réseau de neurones biologiquement plausible reproduisant les *patterns* d'activités décrits ci-après se trouvent dans le papier de Bogacz et al. (2001b).

Les notations sont similaires à celles utilisées dans l'article de Bogacz & Brown (2003b). La couche d'entrée contient n neurones et la couche de sortie contient m neurones de nouveauté, où $n = m$ dans nos simulations. Les deux couches sont entièrement connectées l'une à l'autre. Les poids entre un neurone d'entrée j et un neurone de nouveauté i sont notés w_{ij} . Ils sont initialisés aléatoirement entre -1 et 1 au lancement du réseau. L'activité d'un neurone d'entrée j est notée x_j .

Notre modèle Hebbien ne fonctionne pas avec des valeurs binaires (-1, 0 ou 1) pour déterminer l'activité (ou inactivité) des neurones. Dès lors, nous postulons qu'un neurone d'entrée est actif s'il possède une valeur positive. S'il possède une valeur négative, il sera considéré comme inactif.

Pour limiter le nombre de neurones de nouveauté actifs, nous postulons également qu'un neurone de nouveauté i est actif uniquement si le neurone d'entrée j correspondant est également actif. Cette relation correspond à une forte connexion non-modifiable entre les deux. Ainsi, le passage par cette forte connexion est nécessaire pour que le neurone de nouveauté soit actif. Lorsque le modèle reçoit pour la première fois une nouvelle image X , nous partons donc du principe que les neurones de nouveauté i possédant ce type de fortes connexions seront inexorablement à l'état actif tandis que ceux n'en possédant pas seront inactifs. Pour rappel, certains neurones d'entrée j sont inactifs et les neurones de nouveauté i correspondants seront dès lors inactif.

En partant de ces deux postulats, lors de la première présentation d'une image X , le *pattern* d'activités des neurones de nouveauté i est le même que celui des neurones d'entrée j (**Figure 16**). En d'autres termes, nous considérons que la réponse initiale du réseau se traduit comme suit :

$$x_j^X = x_i^X$$

où x^X représente le vecteur de l'activité des neurones pour une image X .

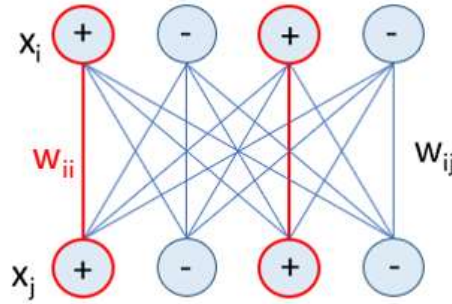


Figure 16 | Représentation schématique de notre réseau Hebbien, où x_j correspond à l'activité des neurones d'entrée pour une image tandis que x_i correspond à l'activité des neurones de nouveauté en réponse à cette image. Les poids sont représentés par les liens w_{ij} entre les neurones et les fortes connexions correspondent aux liens w_{ii} rouges.

Lors de la présentation d'une nouvelle image X pendant l'entraînement, nous ne tenons donc pas compte de la réponse initiale du réseau, puisque nous considérons qu'elle est identique à l'activité des neurones d'entrée i . Ainsi, nous modifions directement l'intégralité des poids des neurones de nouveauté selon la formule suivante :

$$w_{ij} = w_{ij} + \eta x_i x_j$$

où $\eta > 0$ correspond à la constante d'apprentissage, x_i et x_j correspondent respectivement aux éléments qui composent les vecteurs des deux neurones postsynaptique et présynaptique, c'est-à-dire aux caractéristiques de l'image X obtenues à la sortie du RPC⁴.

Lors de la phase d'entraînement, cette modification de poids est implémentée sur 1 *epoch* pour chaque image X du *training set*. Grâce à cet algorithme, nous reproduisons les mécanismes synaptiques de LTP quand x_i et x_j sont positifs, c'est-à-dire que nous renforçons les connexions quand les deux neurones sont activés ensemble (Hebb, 1949). Dans le cas contraire, nous diminuons ces connexions avec les mécanismes de LTD hétéro-synaptiques lorsque x_i est positif (actif) et x_j négatif (inactif) ainsi que de LTD homo-synaptiques lorsque x_i est négatif et x_j est positif.

Lors de la phase de test, nous calculons le potentiel membranaire (h_i) des neurones de nouveauté en tenant compte de l'ensemble des connexions à l'exception des fortes connexions (w_{ii}). En effet, étant donné que nous assumons que ces fortes connexions ne sont pas modifiables, les poids correspondants ne participent pas à l'encodage des caractéristiques de

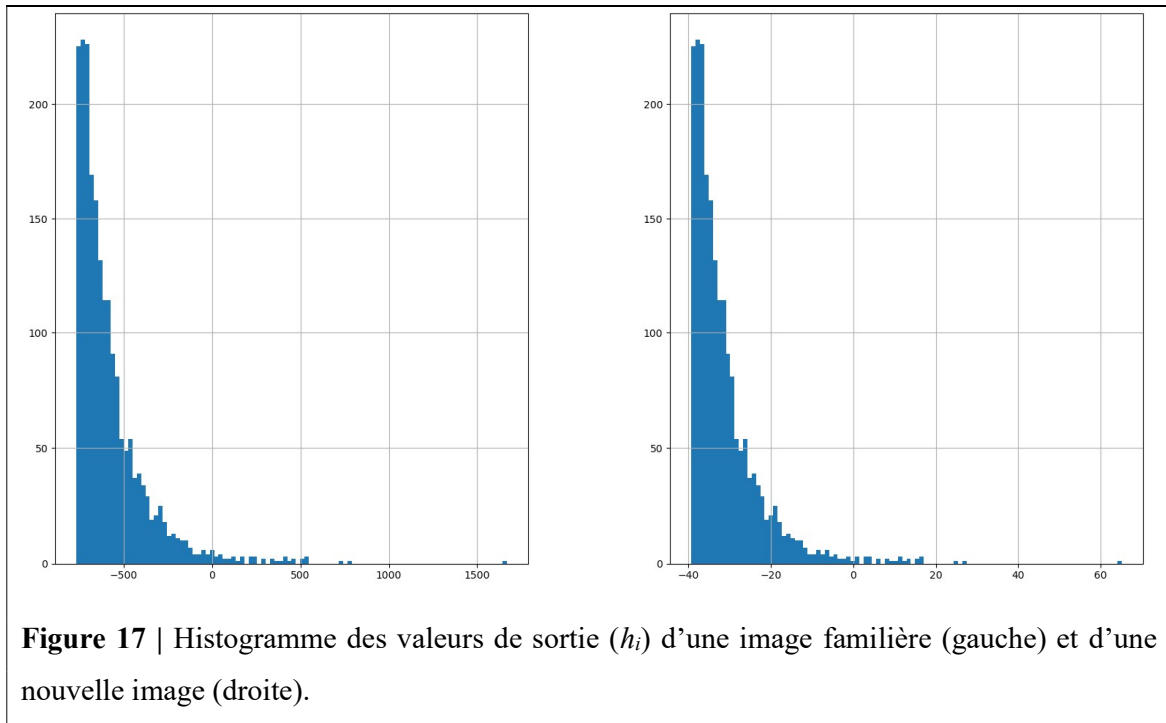
⁴ Rappelons encore une fois que le vecteur x_i est identique au vecteur x_j .

l'image. Nous appliquons la formule suivante :

$$h_i = \sum_{\substack{j=1 \\ j \neq i}}^m w_{ij} x_j, \quad i = 1, \dots, m$$

où x_j correspond au vecteur de caractéristiques de l'image X, produit en sortie du RPC et après normalisation, et w_{ij} correspond aux poids entre les neurones des couches d'entrée et de sortie.

La **Figure 17** représente les histogrammes des valeurs de l'activité d'une image familière par rapport à une nouvelle image.



Ensuite, comme mentionné précédemment, l'apprentissage Hebbien tel que présenté par Bogacz et al. (2001b) postule que les neurones de nouveauté projettent sur un interneurone inhibiteur, qui va inhiber leur activité pour les images familières. Le niveau d'inhibition $d(X)$ est calculé avec la formule suivante :

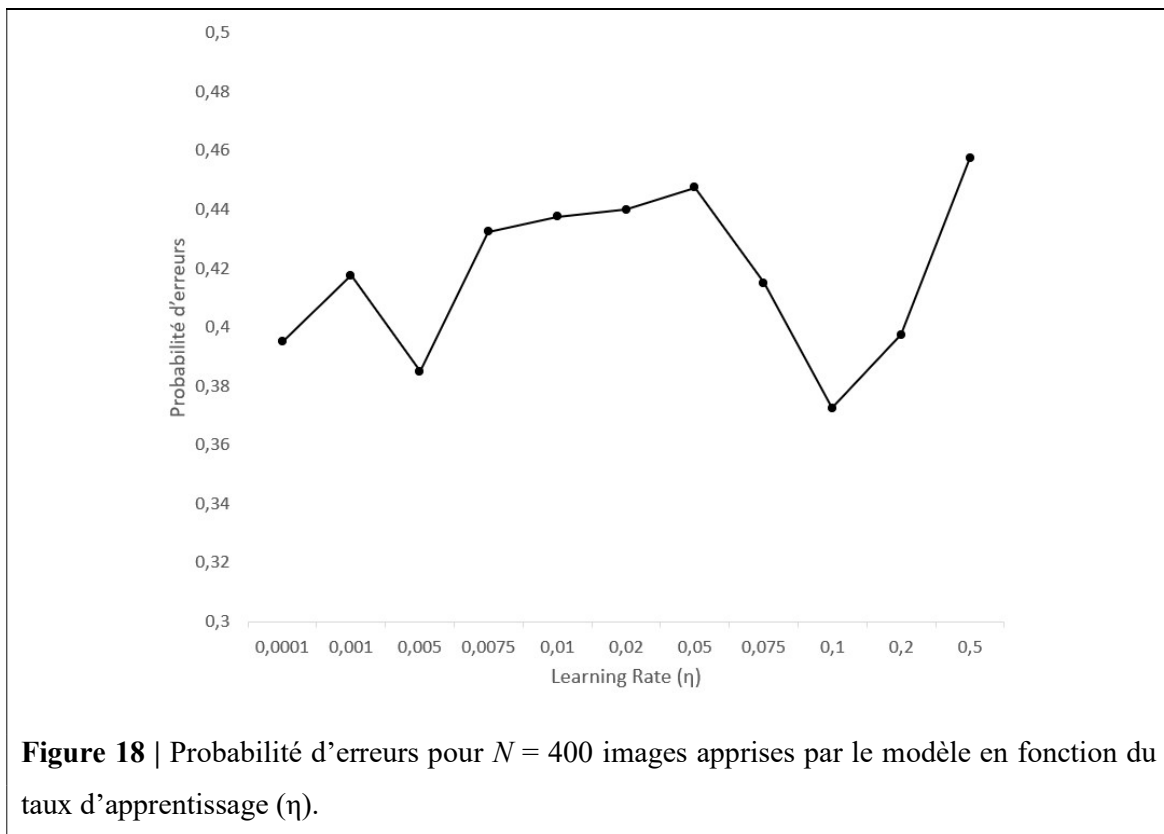
$$d(X) = \sum_{i=1}^m x_i h_i$$

où h_i correspond au potentiel membranaire des neurones de nouveauté et x_i correspond au vecteur de caractéristiques de l'image X obtenues à la sortie du RPC.

Selon le modèle Hebbien, c'est ce niveau d'inhibition qui permet la réduction de l'activité observée dans le cortex périrhinal. Si ce mécanisme n'est pas implémenté dans un modèle comme celui-ci⁵, utilisant des RNA classiques (voir Bogacz et al. (2001b) pour l'architecture d'un modèle Hebbien complet), nous en gardons tout de même les fondements pour notre fonction de décision. Plus précisément, lors d'une tâche de reconnaissance à choix forcés entre une paire d'images (X, Z), où X est une image étudiée à l'entraînement et Z est une nouvelle image, un jugement de familiarité est correct si $d(X) > d(Z)$.

2.3. Paramètres de base du réseau

Les propriétés initiales du module de mémoire sont les suivantes : $n = 2048$ neurones, $m = 2048$ neurones de nouveauté et $\eta = 0,1$. Le choix de la constante d'apprentissage (η) a été fixé arbitrairement par essai-erreur. Plus précisément, nous avons calculé la probabilité d'erreurs du modèle à une tâche de reconnaissance lorsque le nombre d'items étudiés correspond à $N = 400$. La valeur de η retenue correspond à celle pour laquelle la probabilité d'erreurs est la plus faible. Nous avons fait varier ce taux d'apprentissage entre 0.001 et 0.5 sans constater de changement radical dans les performances du modèle (**Figure 18**).



⁵ On peut dès lors se questionner sur la complétude de notre modélisation, ce point sera abordé dans la discussion.

Les différentes simulations prennent la forme d'un test de RCF tel que réalisé en neuropsychologie clinique. Lors d'une simulation, deux échantillons de N images différentes sont aléatoirement sélectionnés parmi l'entièreté du jeu de données (i.e. *dataset*) utilisé. Le premier échantillon contient les cibles et constitue le *training set*, tandis que le second contient les leurres. Toutes les images du *training set* sont apprises une par une par le modèle, grâce à une mise-à-jour des poids selon la règle d'apprentissage. Lors de la phase de test, le modèle recevra une paire d'images ($X_k, Z_k, k = 1, \dots, N$), où la première est une cible tandis que la seconde un leurre. Ainsi, le modèle recevra N paires d'images et procèdera à un jugement de familiarité selon la fonction de décision du modèle. Tous les scores calculés lors des diverses simulations seront moyennés sur 100 *runs* du modèle.

3. Simulations

Maintenant que nous avons exposé le fonctionnement de notre modèle ainsi que défini ses paramètres, nous pouvons présenter les différentes simulations réalisées dans le cadre de ce mémoire. Elles sont au nombre de quatre et explorent chacune un *benchmark* de la familiarité.

3.1. Capacité de mémoire

La première simulation réalisée dans le cadre de ce travail est une reproduction de l'expérience psychologique de Standing (1973). Elle a pour objectif d'investiguer et de comparer la capacité mnésique du modèle à l'issue de tests de RCF entre deux images. Dans cette simulation, le nombre d'images apprises pendant l'entraînement augmente au fur et à mesure des essais.

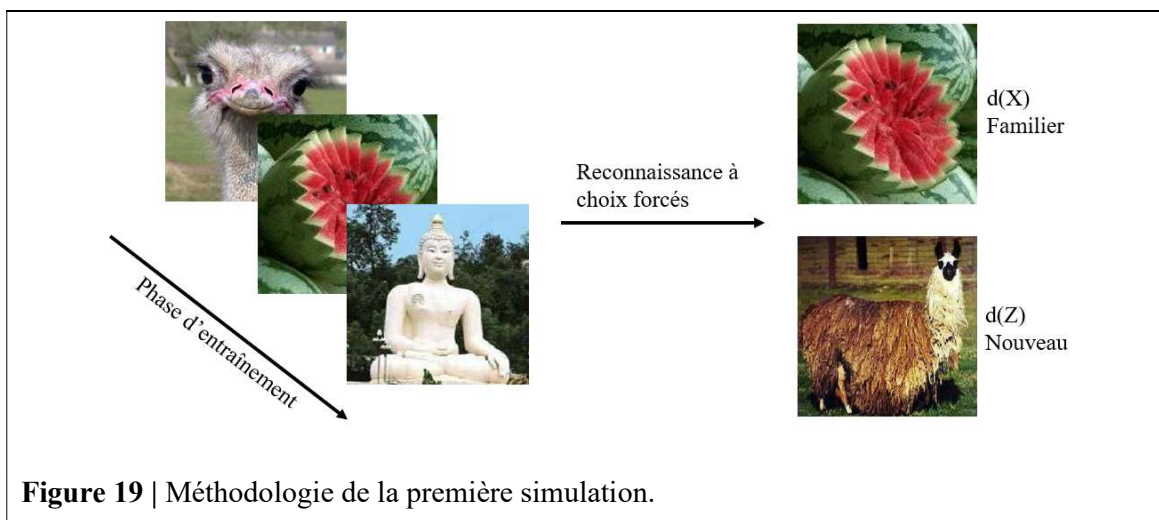
3.1.1. Dataset

Le *dataset* utilisé a été téléchargé sur Kaggle et provient de la base de données d'images « Caltech 256 » (Griffin et al., 2007). Ce *dataset* contient plus de 30 000 photographies d'images naturelles, réparties sur 257 catégories d'objets. Les catégories sont très variées, allant de crapauds à des balles de golf, et contiennent en moyenne 119 images. Chaque catégorie contient au minimum 80 images.

3.1.2. Méthodologie

Lors de cette simulation (**Figure 19**), le modèle est entraîné sur un sous-ensemble d'images de taille N , qui sont apprises une par une grâce à une modification des poids selon la règle d'apprentissage utilisée. Chaque image est présentée une seule fois au modèle. Ensuite, lors de la phase de test, N paires d'images de catégories différentes sont présentées au modèle.

La première est une cible tandis que la seconde est une nouvelle image sélectionnée aléatoirement dans le *dataset*. Le modèle procède à un jugement de familiarité, enregistre sa réponse et l'action est répétée pour les N paires d'images.



La probabilité d'erreurs est ensuite calculée pour l'ensemble de la tâche, avant d'être moyennée sur 100 simulations afin d'obtenir la moyenne et l'écart-type. Chaque simulation est réalisée avec des *training* et *testing set* différents. Le nombre d'images retenues en mémoire après l'apprentissage est calculé avec la même formule que dans l'expérience originale de Standing (1973) :

$$N_{ret} = N(1 - 2P_{er})$$

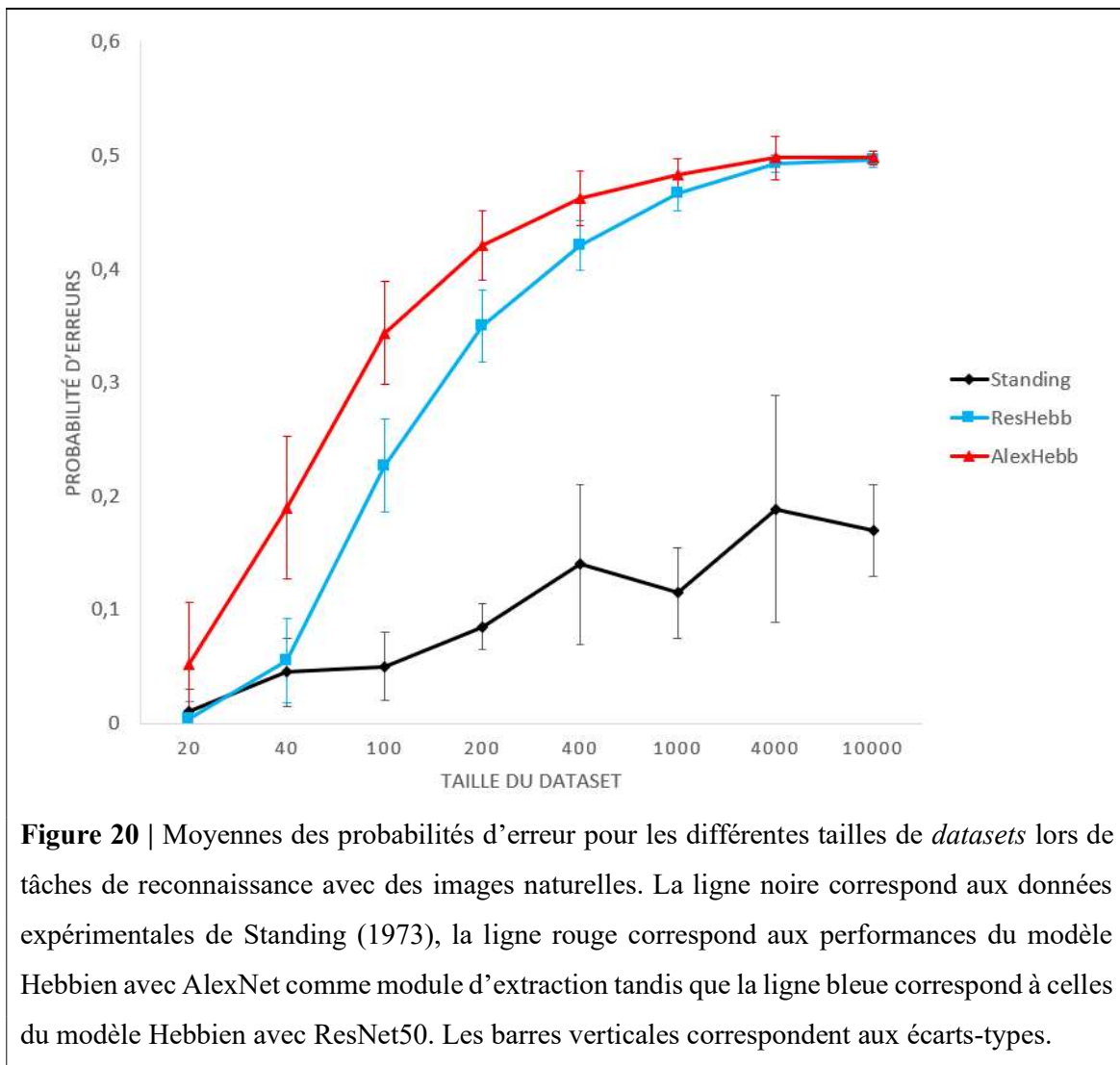
où N est le nombre d'images présentées pendant l'entraînement et P_{er} est la probabilité d'erreurs.

Cette procédure a été reproduite plusieurs fois en augmentant la taille N du *dataset*. Les simulations ont été réalisées pour les nombres d'images suivants : 20, 40, 100, 200, 400, 1000, 4000, 10000.

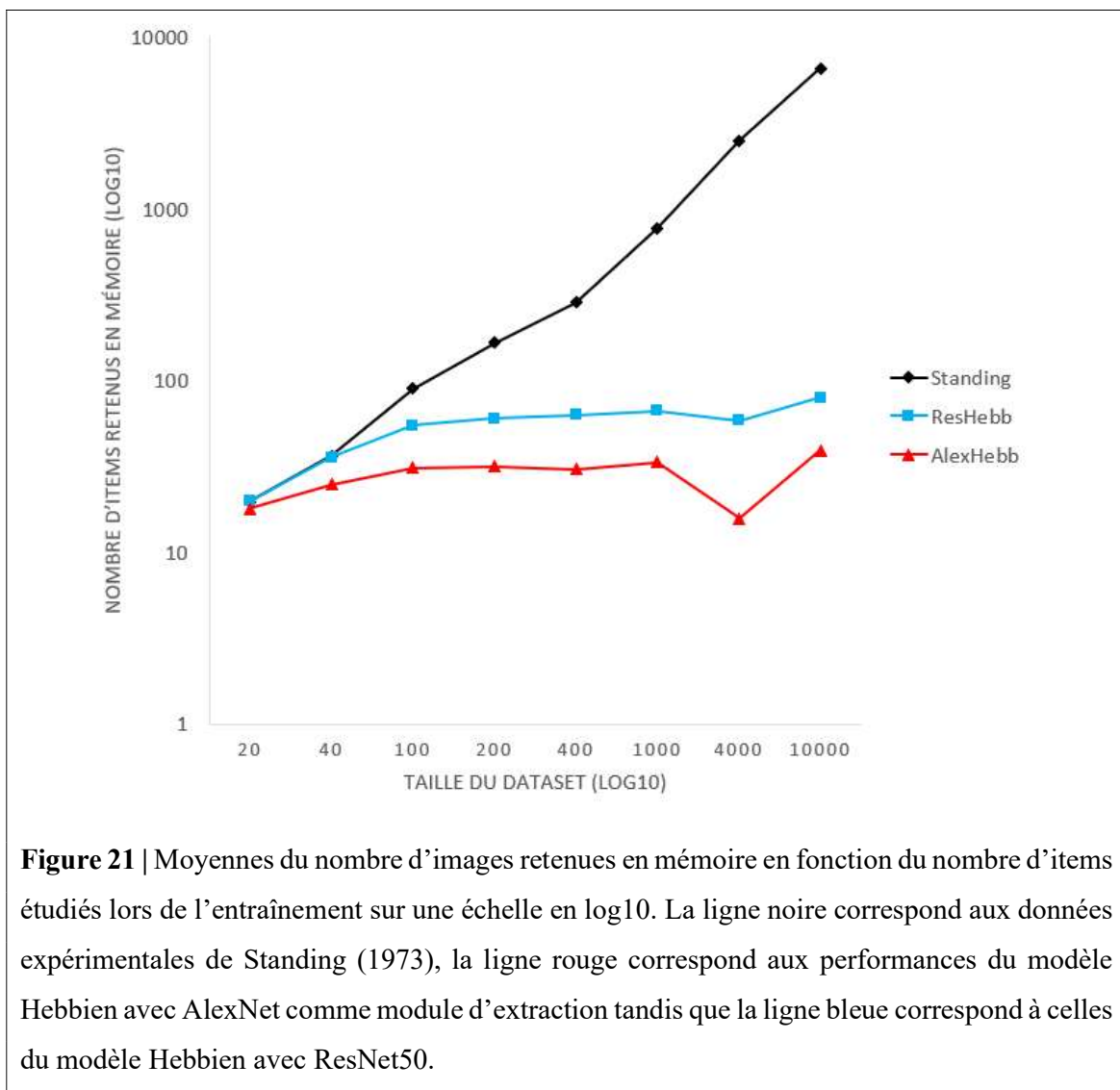
3.1.3. Résultats

Le premier constat intéressant sur le graphique ci-dessous est que la probabilité d'erreurs croît au fur et à mesure de l'augmentation du nombre d'images étudiées par le modèle (**Figure 20**). Néanmoins, les données obtenues avec le module de mémoire Hebbien ne correspondent pas très bien à celles de Standing (1973), représentées par les lignes noires. Cette différence est particulièrement notable lorsque $N > 100$. Par ailleurs, le taux d'erreurs semble atteindre un pallier lors $N > 4000$, avec une probabilité d'erreurs d'environ 0.5, rappelant

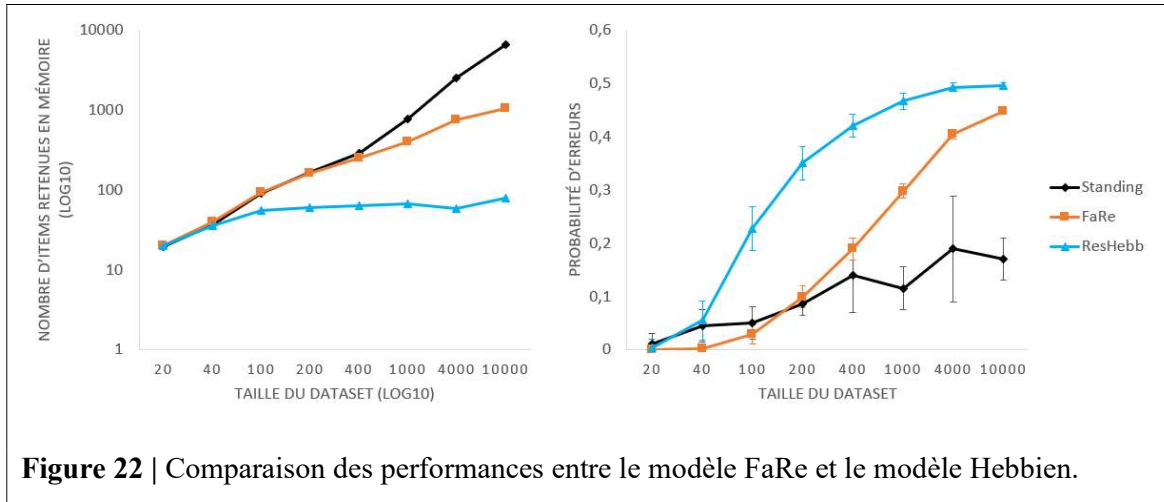
fortement la probabilité du hasard. Notons également que la modification de la *learning rate* ne semble pas améliorer les performances du modèle Hebbien. Néanmoins, il est intéressant de constater que les performances du modèle avec ResNet50 comme module d'extraction de caractéristique sont supérieures par rapport à AlexNet. C'est particulièrement le cas lorsque $N \leq 40$, pour lequel nous obtenons des résultats qui correspondent bien aux données expérimentales.



La **Figure 21** nous confirme en outre les pauvres performances du modèle pour retenir les images en mémoire lorsque $N \geq 100$. De plus, ce nombre ne semble pas augmenter au-delà de 100 images retenues en mémoire pour le modèle avec ResNet50. Par ailleurs, moins de 50 images sont retenues en mémoire pour le modèle avec AlexNet. Notons encore une fois que les courbes collent relativement bien avec les données expérimentales lors que la taille du *dataset* correspond à $N = 20$ et $N = 40$.



La **Figure 22** nous permet d’établir une comparaison entre les performances du modèle FaRe – implémenté sur Python sur base de l’article de Kazanovich et Borisyuk (2021) – et celles de notre modèle Hebbien, avec ResNet50 pour module d’extraction. Il est manifeste que le modèle FaRe surpasse le nôtre quant à ses performances, tant pour le nombre d’images retenues en mémoire que pour la probabilité d’erreurs. Nous pouvons également constater que la courbe orange obtenue pour le modèle FaRe se rapproche sensiblement de la courbe noire qui correspond à l’expérience de Standing (1973). Cependant, les performances du modèle FaRe s’écartent fortement des données expérimentales lorsque $N > 1000$. Dans les simulations des auteurs (Kazanovich et Borisyuk, 2021), cette surestimation des performances n’était présente qu’à partir de $N > 4000$. Lorsque la taille du *dataset* est de $N = 40$, les résultats de notre modèle correspondent davantage à ceux de Standing (1973), contrairement au modèle FaRe qui est « trop » performant dans ces conditions.



3.2. Effets de récence et de primauté

La seconde simulation a pour objectif d’investiguer la présence ou l’absence d’un effet de récence et de primauté lors de tests de RCF, en fonction de l’augmentation du nombre d’images insérées dans le modèle. Pour ce faire, nous avons comparé les probabilités d’erreurs moyennes sur les premières images avec lesquelles le modèle a été entraîné par rapport à celles des dernières images étudiées. Les configurations méthodologiques sont similaires à celles de la simulation précédente ; l’expérience a été réalisée sur le modèle FaRe original, ainsi que sur le modèle Hebbien avec ResNet50 comme module d’extraction de caractéristiques.

3.2.1. Méthodologie

Le *dataset* utilisé est identique à la Simulation 1 (Griffin et al., 2007). Lors de cette simulation, le modèle est entraîné sur un sous-ensemble d’images de taille $N = 30$, puis $N = 200$, qui sont apprises une par une grâce à une modification des poids selon la règle d’apprentissage utilisée. Chaque image est présentée une seule fois au modèle. Ensuite, lors de la phase de test, N paires d’images – 1 ancienne et 1 nouvelle – de catégories différentes sont présentées au modèle (voir **Figure 19**).

L’ordre de présentation des paires d’images est identique à celui de l’entraînement. On calcule la probabilité que le modèle commette une erreur pour une paire d’images donnée sur 100 simulations et ce, pour chaque paire d’images lors du test. S’il y a un phénomène de récence, l’erreur moyenne pour les images étudiées au début de l’entraînement est censée être plus élevée que pour les images récentes, étudiées à la fin de l’entraînement. Un effet de primauté se marquera au contraire par un score plus faible au début, c’est-à-dire pour les premières images du *dataset* par rapport au reste des images.

3.2.2. Résultats

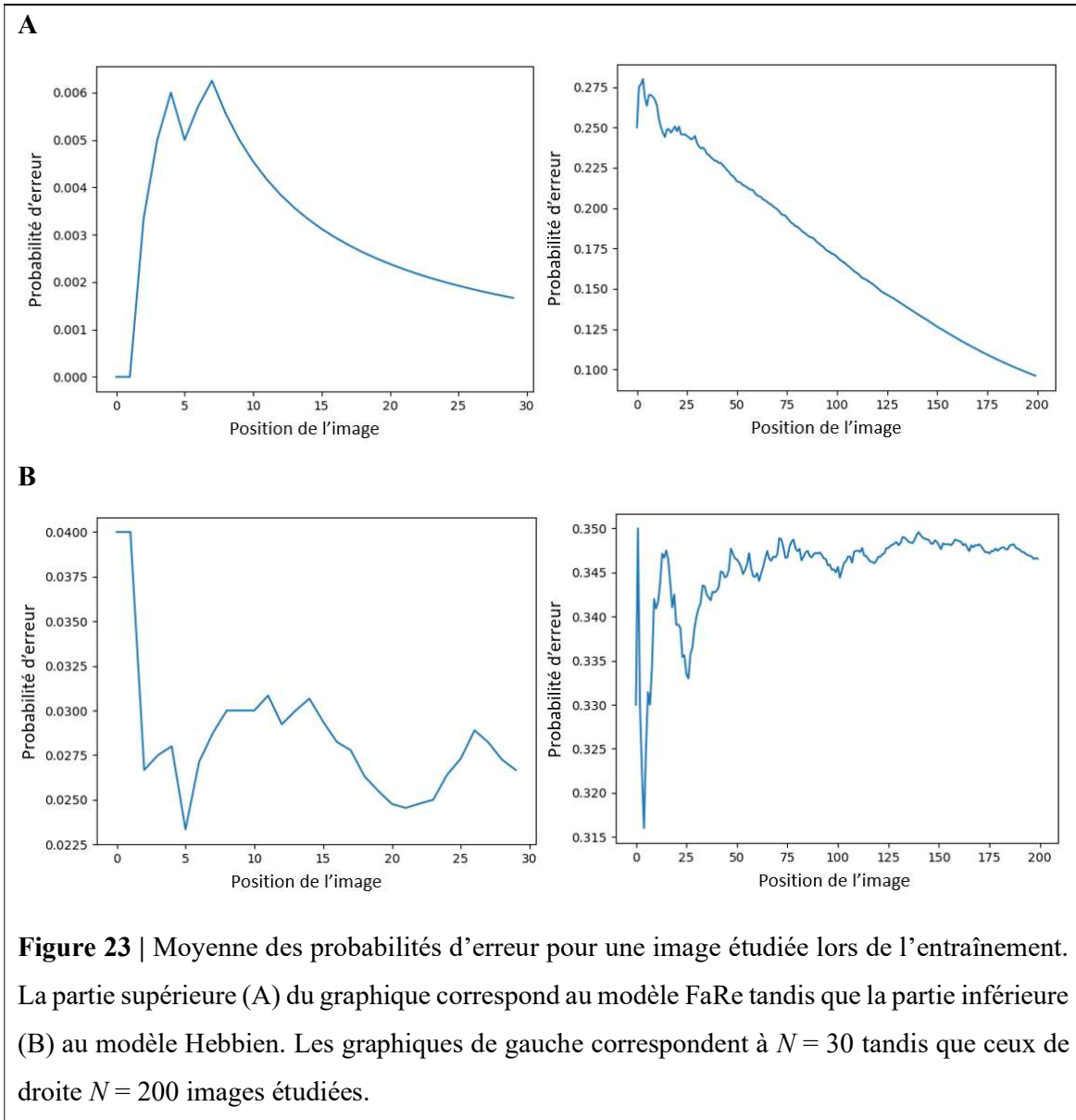


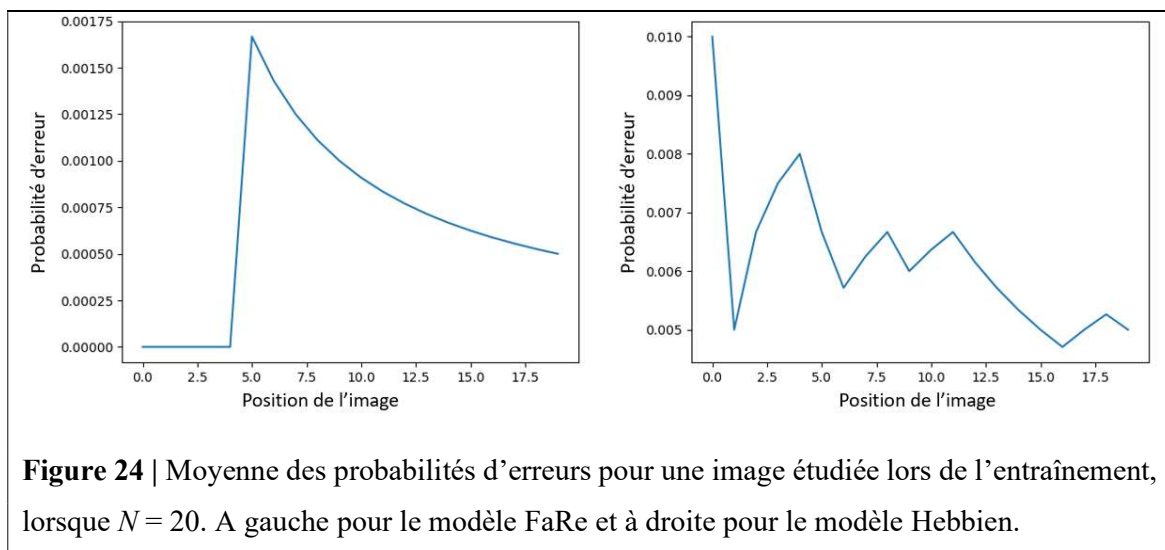
Figure 23 | Moyenne des probabilités d'erreur pour une image étudiée lors de l'entraînement. La partie supérieure (A) du graphique correspond au modèle FaRe tandis que la partie inférieure (B) au modèle Hebbien. Les graphiques de gauche correspondent à $N = 30$ tandis que ceux de droite $N = 200$ images étudiées.

Commençons par l'analyse des performances du modèle FaRe (**Figure 23A**). Rappelons d'abord que tous les résultats ont été calculés sur 100 simulations du modèle afin d'obtenir une moyenne des scores. Dans un premier temps, nous pouvons observer une courbe qui descend progressivement au cours du test. Cela signifie que la probabilité que le modèle produise une erreur de familiarité pour la 5^{ème} paire d'images présentée lors du test est supérieure à celle pour la 25^{ème} paire d'images par exemple. En d'autres termes, les images qui sont étudiées à la fin de la phase d'entraînement sont plus souvent reconnues que les images qui sont étudiées au début de l'entraînement. Cela correspond à un phénomène de récence, tant lorsque $N = 30$ que lorsque $N = 200$.

Dans le graphique supérieur droit, nous pouvons par ailleurs observer que le modèle ne réalise aucune erreur pour la première image, suggérant un effet de primauté. Cet effet n'apparaît plus lorsqu'on augmente la taille du *dataset* ($N = 200$). Bien qu'il soit probable que ce résultat corresponde à du bruit modélisé par le modèle, rappelons qu'il a été calculé à partir de 100 simulations pour lesquelles le modèle a toujours retrouvé correctement la première image. Nous avons testé la présence de cet effet en réduisant le nombre d'items ($N = 20$) et avons constaté que cet effet de primauté s'étend à plus d'images (**Figure 24**).

Concernant l'analyse des graphiques obtenus avec le modèle Hebbien (**Figure 23B**). Nous observons sur le graphique de gauche ($N = 30$) que le tracé n'est pas aussi lisse qu'avec le modèle FaRe. Par exemple notre modèle fait en moyenne moins d'erreurs pour la 5^{ème} paire d'images que pour la dernière, ce qui va à l'encontre d'un effet de récence. Néanmoins, une tendance générale vers le bas se dégage, le modèle faisant en moyenne moins d'erreurs pour la dernière paire que pour la première. Si nous réduisons encore la taille du *training set* ($N = 20$), cet effet de récence est davantage marqué (**Figure 24**). Lorsqu'on augmente le nombre d'items étudiés pendant la phase d'entraînement ($N = 200$), aucun effet de récence n'est observé.

Aucun effet de primauté n'est observé lorsque $N = 30$. Dans le graphique inférieur droit de la **Figure 23B**, on pourrait naïvement imaginer un effet de primauté étant donné que la courbe monte légèrement lorsque $N = 200$. En effet, le modèle commet en moyenne moins d'erreurs pour la première image que pour la dernière. Toutefois la présence de fortes fluctuations en début de simulation, observables via les pics qui se forment lors du passage de la première image à la deuxième, et ainsi de suite. Ces fluctuations sont trop importantes pour permettre de conclure à un véritable effet de primauté.



3.3. Similarité visuelle

La quatrième simulation a pour objectif d'explorer les performances des modèles lorsque les distracteurs appartiennent à la même catégorie que les cibles, en fonction du format de l'épreuve de reconnaissance administrée. En effet, plus la similarité entre les leurres et les cibles augmente et davantage les performances du modèle devraient chuter (Hintzman et al., 1992). Afin de modéliser la similarité entre les leurres et les cibles, nous avons utilisé des images provenant de la même catégorie sémantique lors des simulations, autrement dit des images similaires. Cette simulation a été réalisée sur le modèle FaRe ainsi que sur le modèle Hebbien.

3.3.1. Dataset

Le *dataset* utilisé a été téléchargé sur Kaggle et s'intitule « Cat Dataset » (W. Zhang et al., 2008). Il contient environ 10 000 images de chats réparties dans 7 dossiers. Pour chaque image, les auteurs ont annoté la tête du chat avec 9 points de repère. Les points de repère correspondent aux deux yeux, à la bouche et aux 6 angles qui constituent les deux oreilles du chat.

3.3.2. Méthodologie

La méthodologie est similaire celle réalisée lors de la simulation 1, à l'exception de la taille du *training set* qui contient $N = 40$ images de chats. Lors de l'entraînement, chaque image est présentée une seule fois au modèle. Le *testing set* contient 40 paires d'images, dont un chat familier et un nouveau chat pour chaque paire (**Figure 25**). La probabilité d'erreurs moyenne ainsi que l'écart-type sont calculés sur 100 simulations, avec des *training* et *testing set* différents pour chacune d'elles.

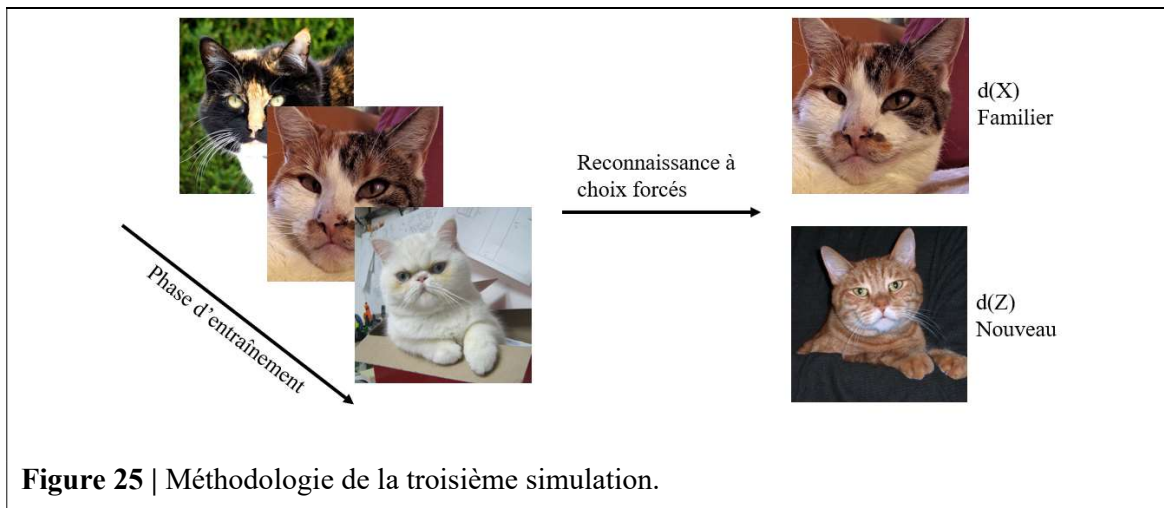


Figure 25 | Méthodologie de la troisième simulation.

3.3.3. Analyses statistiques

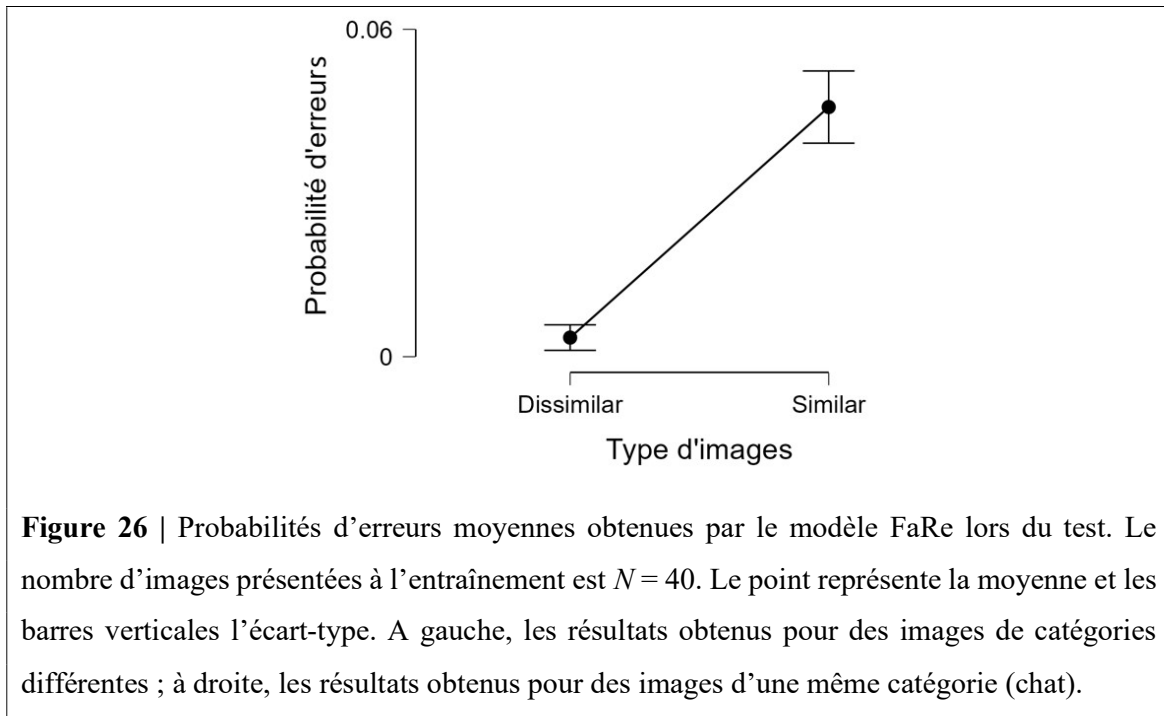
Les analyses statistiques ont été réalisées sur le logiciel JASP (2020). Un seuil statistique de significativité correspondant à $\alpha = 0.05$ a été utilisé. Nous avons comparé les résultats avec ceux obtenus pour la même simulation réalisée sur des images de catégories différentes (voir **Figure 19**). La normalité des distributions a été préalablement éprouvée via un test de normalité de *Shapiro-Wilk*. Les moyennes dans les deux groupes ont ensuite été comparées à l'aide d'un test non-paramétrique *U de Mann-Whitney*, en raison de la violation de l'hypothèse de normalité. Une corrélation bisériale de rang a finalement été effectuée pour obtenir les tailles d'effet.

3.3.4. Résultats

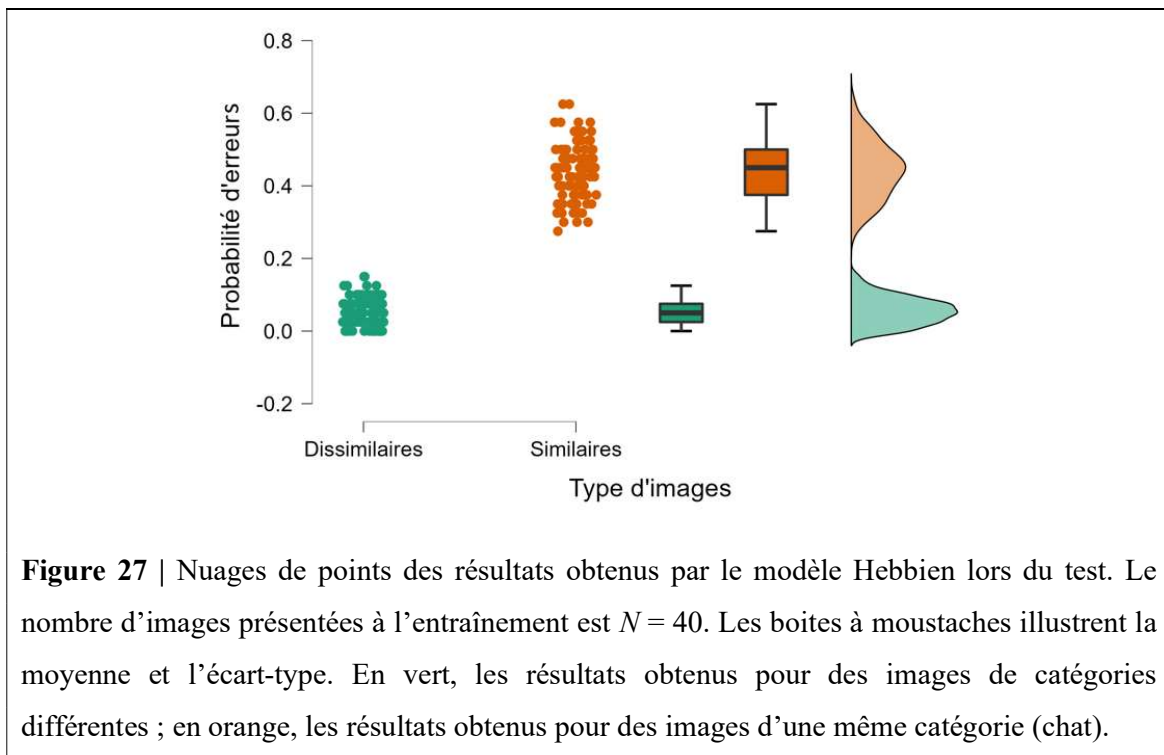
La **Figure 26** correspond aux moyennes et écarts-types des probabilités d'erreurs obtenus lors du test avec le modèle FaRe ; la **Figure 27** correspond aux nuages de points et moyennes des scores obtenus lors du test avec le modèle Hebbien. Le choix de réaliser deux graphiques différents selon le modèle a été effectué pour des raisons de compréhension secondaire à une clarté visuelle optimale des résultats.

Concernant les résultats du modèle FaRe, le test de normalité nous a amené à rejeter l'hypothèse de la normalité de la distribution des données, tant pour les images similaires ($W = 0.920, p < .001$) que pour les images de catégories différentes ($W = 0.334, p < .001$). Le test *U de Mann-Whitney* a mis en évidence une différence significative ($W = 1175.50, p < .001$) entre les moyennes des probabilités d'erreur pour les deux groupes, avec une taille d'effet $r = -0.7$. Le graphique ci-dessous (**Figure 26**) illustre cette différence significative entre les groupes.

Il est toutefois intéressant de constater que les écarts-types sont plus grands lorsque les images sont similaires ($\mu = 0.046, \sigma = 0.033$). Par ailleurs, la moyenne de la probabilité d'erreurs pour les images de catégories différentes est très proche de 0 ($\mu = 0.004, \sigma = 0.012$). D'un point de vue qualitatif et à l'échelle des performances du modèle, il serait tentant de négliger cette différence de performances, ou tout du moins d'éviter de l'attribuer obligatoirement à un effet de la similarité entre les images.



Concernant les résultats du modèle Hebbien, le test de normalité nous a amené à rejeter l'hypothèse de la normalité de la distribution des données pour les images de catégories différentes ($W = 0.939, p < .001$). Le test U de *Mann-Whitney* a mis en évidence une différence significative ($W = 0, p < .001$) entre les moyennes des probabilités d'erreur pour les deux groupes, avec une taille d'effet $r = -1$.



Le graphique précédent (**Figure 27**) illustre clairement cette différence significative entre la probabilité d'erreurs moyenne pour les images de catégories différentes ($\mu = 0.055$, $\sigma = 0.037$) et les images similaires ($\mu = 0.441$, $\sigma = 0.077$). Sur ce schéma, il est aisé de se rendre compte que les performances du modèle Hebbien chutent drastiquement lorsqu'il est confronté à des images d'une même catégorie. Notons que lorsque les images sont différentes, le modèle parvient sans peine à correctement discriminer les nouvelles des anciennes. Par ailleurs, la probabilité d'erreurs semble plus aléatoire pour les images similaires, en témoigne un écart-type supérieur comparé aux images différentes.

3.4. Orientation de la cible

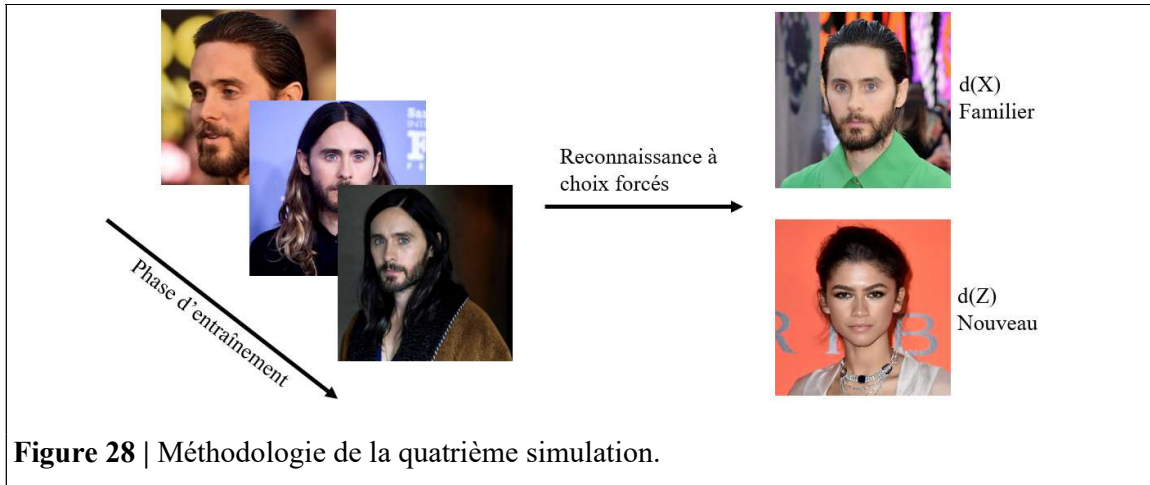
La dernière simulation a pour objectif de reproduire un phénomène observé dans les travaux de Ji-An et al. (2022) sur un modèle utilisant des neurones complexes. Plus précisément, nous avons exploré les performances du modèle lors d'un test de RCF durant lequel les cibles – ici des visages humains - précédemment étudiées sont présentées dans des positions différentes ou avec des caractéristiques physiques différentes (pilosité faciale, l'âge de la personne, ...).

3.4.1. Dataset

Le *dataset* utilisé est un large ensemble de visages intitulé « VGGFace » (Parkhi et al., 2015), accessible en open access et conçu pour la recherche sur la reconnaissance faciale. Il contient environ 2.6 millions de photos de visages appartenant à 2622 personnalités plus ou moins connues, tels que des acteurs ou encore des politiciens. Pour chaque personnalité, le *dataset* comporte en moyenne de 362,6 photos, toutes présentant des variations intrinsèques telles que la position, l'âge, la pilosité faciale mais également l'arrière-plan.

3.4.2. Méthodologie

La méthodologie utilisée est similaire aux précédentes. A l'entraînement, nous présentons au modèle une seule fois N photos de célébrités. Lors de la phase de test, nous présentons au modèle N paires d'images. La première étant une photo d'un visage précédemment étudié avec des caractéristiques différentes tandis que la seconde photo est un nouveau visage jamais présenté auparavant (**Figure 28**). La probabilité d'erreurs moyenne ainsi que l'écart-type sont calculés sur 100 simulations, avec des *training* et *testing set* différents pour chacune d'elles. Les simulations ont été réalisées pour un nombre croissant d'images : 10, 20, 30, 40, 50.



3.4.3. Résultats

Les résultats sont présentés dans le **Tableau 1**. Il reprend les probabilités d'erreurs moyennées sur 100 simulations, réalisées avec le modèle FaRe et le modèle Hebbien, pour différents nombres de photos de visages appris pendant l'entraînement.

Tableau 1 | Moyennes et écarts-types des probabilités d'erreurs obtenus lors d'une tâche de RCF.

Training Set	Modèle FaRe		Modèle Hebbien	
	Moyenne	Ecart-type	Moyenne	Ecart-type
10	0.182	0.120	0.216	0.138
20	0.156	0.092	0.213	0.104
30	0.153	0.063	0.194	0.077
40	0.147	0.058	0.195	0.052
50	0.133	0.050	0.205	0.061

Encore une fois, le modèle FaRe est plus performant que le modèle Hebbien pour l'ensemble des tailles de *training set* proposées. Néanmoins, l'un comme l'autre semble correctement reconnaître des visages familiers présentés dans différents contextes avec une précision supérieure à 80% pour le modèle FaRe et supérieure à 75% pour le modèle Hebbien. Étonnement, les modèles performant moins bien lorsque la taille du *training set* est minimale ($N = 10$). Néanmoins, les performances globales semblent relativement stables au fil des simulations.

PARTIE III : DISCUSSION

La partie introductive de ce mémoire a fait transparaître la complexité des mécanismes de familiarité chez l'homme. Aujourd'hui, les progrès dans le domaine de l'IA nous permettent de clarifier les théories et d'affiner la compréhension de la cognition. L'objectif du présent travail est de fournir de nouvelles données dans l'équation qu'est la familiarité, afin d'en éclaircir un tant soit peu les particularités. Pour ce faire, nous avons implémenté un modèle connexionniste, inspiré des travaux de nos prédécesseurs (Bogacz et al., 2001b ; Kazanovich & Borisjuk, 2021), modèle qui pourrait imiter artificiellement le SdF et ainsi reproduire plusieurs résultats expérimentaux liés à la reconnaissance visuelle. Avec ce modèle, consistant en un réseau convolutif couplé à un apprentissage Hebbien, nous avons modélisé quatre situations faisant appel au processus de familiarité. Ces situations sont discutées individuellement dans cette dernière partie. Par ailleurs, nous comparons également les performances de notre réseau avec celles du modèle FaRe, précédemment proposé par Kazanovich & Borisjuk en 2021.

Premièrement, nous avons reproduit l'expérience réalisée par Standing en 1973, dans laquelle les participants ont dû réaliser un jugement de familiarité sur un très grand nombre d'images. Notre réseau semble moins performant que le modèle FaRe, mais il a été capable de passer une tâche de RCF sur 40 images avec une précision de 94,5%. En revanche, ses réponses deviennent aléatoires lorsque le nombre d'items appris augmente. Les capacités du module d'extraction améliorent néanmoins les performances du modèle.

Deuxièmement, un effet de récence semble aider notre modèle à discriminer correctement une ancienne d'une nouvelle image : il commet moins d'erreurs pour les images étudiées à la fin de l'entraînement que pour celles étudiées au début. Cet effet est d'autant plus marqué que le nombre d'items appris est faible. Lorsqu'il augmente, l'effet de récence disparaît. Aucun effet de primauté n'a été constaté dans notre modèle, contrairement au modèle FaRe.

Troisièmement, nous avons testé l'effet de la similarité entre les images sur les modèles, en évaluant leurs performances lors d'une tâche où les cibles et leurres appartiennent à la même catégorie sémantique. Nous avons observé des différences significatives dans les performances des deux modèles.

Finalement, nous avons réalisé un test de RCF dans lequel le modèle s'est entraîné sur des photos de célébrités. Il a ensuite dû procéder à un jugement de familiarité entre la photo d'une nouvelle personne et une photo prise sous un autre angle de la célébrité vue précédemment. Les modèles parviennent à reconnaître les photos des célébrités avec une précision d'environ 80% pour le modèle FaRe et 75% pour le modèle Hebbien.

1. Capacité de stockage

Il a déjà été démontré précédemment qu'un simple module de mémoire anti-Hebbien est capable d'atteindre une capacité de stockage allant jusqu'à n^2 , où n correspond au nombre d'unités dans le réseau (Androulidakis et al., 2008 ; Bogacz & Brown, 2003a ; Kazanovich & Borisjuk, 2021). Nos simulations avec le modèle FaRe confirme à nouveau cette hypothèse. En effet, malgré un nombre très important d'items appris lors de l'entraînement (10000 images), la probabilité d'erreurs dans le modèle FaRe n'atteint pas 0.5. Ces résultats correspondent aux données expérimentales (Standing, 1973), suggérant une capacité de stockage quasiment illimitée pour la reconnaissance basée sur la familiarité. En outre, sur l'échelle logarithmique, nous avons constaté une relation qui rappelle une loi de puissance (i.e. *power law*) entre le nombre d'items étudiés et le nombre d'items retenus : le nombre d'images retenues en mémoire augmente proportionnellement à la taille du *dataset*. Ces points sont déjà vivement discutés dans l'article original du modèle FaRe (voir Kazanovich et Borisjuk, 2021).

Analysons plutôt les résultats obtenus avec notre modèle Hebbien. Lorsque le nombre d'images étudiées est faible, dans 94% des cas, le modèle arrive à distinguer les images familières des nouvelles. De plus, les résultats obtenus pour des petites tailles de *datasets* sont en parfaite adéquation avec les courbes obtenues par Standing (1973) et surpassent même ceux du modèle FaRe original lorsque $N = 40$. Ces résultats sont encourageants car ils montrent que le modèle Hebbien est capable de procéder à un jugement de familiarité sous certaines conditions, ajoutant ainsi de la crédibilité aux simulations suivantes. Ils suggèrent également que l'apprentissage de la structure globale d'un stimulus serait un mécanisme de familiarité pertinent lors de la reconnaissance, et ce, lorsque la quantité d'informations n'excède pas un certain nombre ; à l'inverse, lorsque le nombre de stimuli est trop élevé, le jugement de familiarité reposerait sur certaines caractéristiques spécifiques d'un stimulus.

Malgré tout, notre modèle présente une faiblesse dans sa capacité mnésique et ne parvient pas à réaliser un jugement de familiarité lorsque la taille du *dataset* augmente drastiquement. En outre, la **Figure 21** ne rend pas compte de la loi de puissance entre le nombre d'images retenues et la taille du *training set*, observée dans la littérature. L'ensemble de ces résultats vont à l'encontre d'une capacité de stockage presque illimitée pour la reconnaissance sur base du SdF. Demeure alors la question de comprendre pour quelles raisons la capacité de stockage du modèle Hebbien n'atteint pas celle du modèle anti-Hebbien.

1.1. Corrélations entre les entrées

Nos résultats contredisent ceux présentés dans l'article de Bogacz et al. (2001b), qui montraient une grande capacité de stockage pour le modèle Hebbien. De plus, en comparant leurs résultats avec un modèle anti-Hebbien, les auteurs ont d'abord observé que la capacité de stockage du modèle Hebbien devrait être du même ordre que celle de son homologue anti-Hebbien (Bogacz & Brown, 2003b). Cependant, cette première comparaison a été effectuée lorsque les composantes du vecteur d'entrées ne présentaient aucune corrélation entre elles. Lorsque ces corrélations sont insérées dans les *patterns* d'entrées, les performances du modèle Hebbien sont significativement inférieures à celle du modèle anti-Hebbien, plus robuste dans ces situations. Selon les auteurs (Bogacz & Brown, 2003b), cela s'expliquerait par le fait que le modèle anti-Hebbien apprend les éléments qui sont propres à un stimulus particulier tandis que le modèle Hebbien apprendrait plutôt les caractéristiques communes à plusieurs stimuli. Ainsi, lorsque les *patterns* d'entrées sont corrélés, l'apprentissage anti-Hebbien supprimerait automatiquement ces corrélations et encoderait uniquement les composantes non-corrélées d'un vecteur d'entrées, augmentant ainsi la capacité de stockage du modèle (Tyulmankov et al., 2022).

Kazanovich et Borisjuk (2021) ont évalué la corrélation entre les caractéristiques des images, c'est-à-dire le vecteur de sortie du module d'extraction. Ils ont constaté la présence de ces corrélations, qui pouvaient par ailleurs être très élevées pour certaines images. Cela expliquerait les faibles performances de notre modèle sur de larges ensembles d'images. Pour éprouver cette hypothèse, il faudrait vérifier la présence de corrélation entre les entrées du réseau après être passé dans ResNet50. Afin de nous assurer que notre module Hebbien a été correctement implémenté, nous pourrions à l'avenir tester ses performances avec des *patterns* d'entrées qui ne présentent aucune corrélation entre eux.

1.2. Taille du module de mémoire

Bogacz & Brown (2003b) ont également montré que la taille du réseau a un impact sur l'influence des corrélations : lorsque le réseau est constitué de beaucoup de neurones, les corrélations entre l'activité des entrées exercent une influence plus délétère sur la capacité mnésique du modèle Hebbien. Par exemple, passer d'un réseau avec 100 neurones de nouveauté à 300 neurones réduit de moitié les performances du réseau (Bogacz & Brown, 2003b). A l'inverse, l'augmentation de la taille du réseau exercerait une influence positive sur le modèle anti-Hebbien en amenuisant l'impact des corrélations.

Pour rappel, notre réseau est constitué de 2048 neurones de nouveauté contre 4096 neurones pour le modèle FaRe. Dans notre cas, ce nombre est largement plus élevé que celui proposé dans les articles initiaux (Bogacz et al., 2001b ; Bogacz & Brown, 2003b), ce qui pourrait expliquer les résultats de nos simulations. Varier la taille du module de mémoire semble être une piste intéressante. Kazanovich et Borisjuk (2021) n'ont pourtant décelé aucune modification des performances du modèle FaRe, que ce soit en augmentant ou diminuant le nombre de neurones du modèle. Nous n'avons pas encore eu l'occasion d'explorer cette hypothèse avec notre modèle Hebbien.

1.3. Oubli catastrophique

Parmi les problèmes récurrents auxquels sont confrontés les modèles connexionnistes, celui de l'oubli catastrophique est peut-être l'un des plus fondamentaux (McCloskey & Cohen, 1989). Sous certaines conditions, l'apprentissage de nouveaux stimuli écrase brutalement tout ce qui a été précédemment appris par le réseau (voir French (1999) pour une revue de la littérature). Cet oubli catastrophique est inhérent à bon nombre de réseaux qui possèdent un seul et unique ensemble de poids dans leur architecture. Paradoxalement, il s'agit là de la caractéristique précise des RNA qui rend possible la généralisation ou leur permet de fonctionner correctement malgré la dégradation des entrées (French, 1999).

Les résultats de la première simulation suggèrent que notre modèle Hebbien est lui aussi sujet au problème de l'oubli catastrophique. A partir d'un certain nombre d'images, le modèle semble saturer et l'apprentissage de nouvelles images efface la trace mnésique des précédentes images. Lorsqu'on regarde l'évolution des valeurs des poids dans le modèle, on peut constater que ces derniers se majorent de façon exponentielle au fur et à mesure de l'apprentissage. Il est donc raisonnable de penser que ces valeurs exponentielles prises par les poids au fur et à mesure de l'apprentissage de nouveaux items sont à l'origine de l'oubli catastrophique dans notre modèle. Afin de minimiser ce problème, nous avons tenté de réduire la constante d'apprentissage (η), espérant ainsi limiter la vitesse de modification des poids. Cependant, nos simulations ont montré que la réduction de η n'a que très peu d'impact sur les performances du modèle (voir **Figure 18**) et n'empêche pas la saturation de sa capacité mnésique. Une autre piste qui permettrait potentiellement de stabiliser le modèle serait de fixer une limite à la valeur que peuvent prendre les poids, limite au-delà de laquelle ils ne pourraient pas aller.

L'oubli graduel est certes un pilier de la cognition humaine, toutefois l'oubli catastrophique n'a pas encore été observé dans le cerveau humain (McCloskey & Cohen, 1989 ; Ratcliff, 1990). Selon McClelland et al. (1995), l'évolution a trouvé le moyen de le surpasser en séparant le système mnésique dans deux régions cérébrales distinctes, l'hippocampe et le néocortex. D'après eux, l'acquisition séquentielle de nouvelles informations conjointement à l'apprentissage graduel de la structure de l'expérience vécue causerait des interférences avec les anciennes informations encodées.

Dans notre modèle Hebbien, lorsqu'une image est présentée à l'entraînement, ses caractéristiques individuelles sont extraites afin d'être encodées dans un modèle de mémoire. Cette opération est ainsi répétée de façon séquentielle pour chaque image du *training set*. Or, comme exposé précédemment, la règle d'apprentissage Hebbienne fonctionnerait via l'apprentissage des caractéristiques communes entre plusieurs images. Donc, plus le nombre de stimuli augmente et plus le modèle décèlerait des structures partagées par un grand nombre d'items. A l'instar de McClelland et al. (1995), nous émettons donc l'hypothèse que l'encodage de la structure globale d'un stimulus tel que réalisé par l'apprentissage Hebbien est incompatible avec l'extraction des caractéristiques individuelles d'une grande quantité d'images. Ceci pourrait donc être à l'origine de l'oubli catastrophique observé dans notre modèle.

Tout comme l'hippocampe prend le rôle de médiateur pour le néocortex, notre modèle aurait besoin d'un mécanisme d'acquisition rapide des nouvelles informations, évitant ainsi les interférences avec les régularités précédemment découvertes (McClelland et al., 1995 ; O'Reilly et al., 2014). En d'autres mots, la familiarité telle que modélisée dans notre réseau aurait besoin de mécanismes de recollection pour fonctionner à l'égal du cerveau humain. Cette hypothèse est renforcée par le constat que le modèle anti-Hebbien est nettement plus robuste face à l'oubli catastrophique puisqu'il fonctionne via l'apprentissage spécifique des caractéristiques extraites et non l'apprentissage global de la structure de l'item. Elle met également en évidence la complémentarité des mécanismes de recollection et de familiarité, lesquels ne devraient plus s'envisager l'un sans l'autre pour rendre compte de toutes les subtilités de la MdR (Kazanovich & Borisyuk, 2021).

2. Rôle des effets de récence et de primauté

Des études (Deese & Kaufman, 1957 ; Murdock, 1962) ont déjà montré que, lors d'une tâche de rappel libre durant laquelle le nombre d'items varie entre 10 et 40, les participants rappellent mieux les premiers et derniers items de la liste. Ces résultats font référence aux effets de primauté et de récence. Selon les modèles postulant une distinction entre différents types de mémoires (Squire, 2004 ; Squire & Zola, 1998 ; Tulving, 1985), les premiers items de la liste sont stockés dans la mémoire à long terme tandis que les derniers sont entreposés dans la mémoire de travail (ou mémoire à court terme). Ainsi, le phénomène de récence devrait être considéré comme une caractéristique de la mémoire à court terme. Une double dissociation devrait donc être observable pour ces deux effets temporels. Pour explorer cette double dissociation, des simulations futures pourraient être envisagées, à l'instar des travaux réalisés par Sougné (2002), dans lesquelles une tâche de mémoire à court terme correspondrait au rappel d'une série d'images tandis qu'une tâche de mémoire à long terme correspondrait à l'association d'une image à sa catégorie sémantique.

Revenons-en aux résultats de nos simulations. Dans le modèle FaRe, nous avons observé que lorsque la taille du *dataset* est faible, l'effet de primauté est présent ; il en va de même pour l'effet de récence. En effet, lorsque $N = 30$, le premier item est systématiquement rappelé et les derniers items de la liste sont également plus fréquemment reconnus. Par ailleurs, ces résultats sont obtenus alors que le modèle présente un système mnésique unique pour les mémoires à court et à long terme ; nous n'avons pas cherché à distinguer la contribution de ces deux mémoires dans notre modélisation de la familiarité. Néanmoins, nous constatons que lorsque la taille du *dataset* est élevée, l'effet de primauté disparaît à l'inverse de l'effet de récence, qui résiste à l'augmentation du nombre d'images apprises par le modèle. Dans le modèle Hebbien, seul un effet de récence est observable. Cet effet est en outre uniquement présent lorsque la taille du *dataset* est très faible.

D'après Tulving (1985), la familiarité correspondrait à une forme de mémoire déclarative et serait donc envisagée comme un mécanisme de mémoire à long terme. Nous aurions donc pu nous attendre au maintien de l'effet de primauté malgré l'augmentation du nombre d'items à l'entraînement. Les résultats de nos simulations montrent pourtant l'inverse ; l'effet de primauté disparaît au contraire de l'effet de récence. En admettant que l'effet de récence serait une caractéristique de la mémoire de travail, nous émettons donc l'hypothèse que la mémoire à court terme contribue également au SdF. Logie (1996) avait déjà associé la

mémoire à long terme avec la mémoire de travail, en postulant que l'information manipulée par cette dernière permettrait l'activation de la trace en mémoire à long terme. Plus récemment, les travaux de Blalock (2015) suggèrent que des représentations robustes en mémoire à long terme peuvent augmenter les performances en mémoire à court terme visuelle pour des stimuli familiers.

S'il est aujourd'hui évident qu'il existe d'importantes connexions entre les mécanismes à court et à long termes dans l'apparition de la familiarité, il serait néanmoins prématuré de conclure que cette dernière surgirait uniquement à partir de la récence d'un événement et de l'implication de la mémoire de travail. Au contraire, il existerait une multitude de sources pouvant contribuer à l'apparition d'un SdF (Bastin et al., 2019). La contribution de la mémoire de travail, de l'effet de récence ou encore de l'effet de primauté ne seraient que quelques-unes des potentielles sources à la base d'un jugement de familiarité.

2.1. Effet de récence

Les résultats du modèle FaRe lors de la deuxième simulation indiquent vraisemblablement un effet de récence. Le modèle commet moins d'erreurs pour les images qui ont été étudiées à la fin de l'entraînement par rapport aux images étudiées au début de l'entraînement. Comme expliqué, en faisant varier le nombre d'images du *training set*, nous constatons la solidité de cet effet, y compris lorsque le modèle est confronté à un grand nombre de stimuli. Ces résultats concordent avec ceux de la littérature, obtenus par Whittlesea (1993). L'auteur postulait en effet qu'une sensation subjective de familiarité vis-à-vis d'un événement peut être modulée selon sa récence. En d'autres termes, un événement qui a été vécu récemment présenterait une plus forte familiarité par rapport à un autre événement antérieur au premier.

Une explication possible à la présence de cet effet réside dans la perte de la trace (i.e. *trace decay*), également responsable de l'oubli catastrophique dans le modèle (Sougné, 2002). En effet, les poids entre les nœuds qui répondaient fortement aux premières images à l'entraînement vont subir beaucoup de modifications au fur et à mesure de l'apprentissage. Comme les poids de l'image précédente sont réutilisés pour l'image suivante, plus on s'enfonce dans le *training set* et moins les poids d'une image donnée subiront des modifications. Au terme de l'entraînement, les poids utilisés à la phase de test auront des valeurs sensiblement plus proches des valeurs de poids des dernières images apprises par le modèle. Ainsi, la diminution des poids pendant l'entraînement correspondrait à la perte de la trace responsable de l'effet de récence.

Dans le modèle Hebbien, cet effet de récence apparaît également lorsque le nombre d'items présents dans le *training set* est peu élevé. Cependant, pour des grandes tailles de *datasets*, le modèle Hebbien semble perdre cette capacité. La saturation du modèle au-delà d'un certain nombre d'images apprises pourrait expliquer ce phénomène. Comme supposé précédemment, le modèle apprendrait plutôt les *patterns* communs aux stimuli. Lorsqu'il est confronté à un grand nombre de stimuli, ces *patterns* sont donc partagés par énormément d'items, y compris les leurres. En conséquent, plus le nombre d'images augmente et plus le modèle tend à commettre des erreurs, masquant ainsi l'effet de récence attendu. Le modèle ne parviendrait donc plus à faire la différence entre un nouvel item et un ancien et procéderait dès lors à un choix aléatoire. Cette hypothèse est cohérente avec l'aspect de la courbe de droite, visualisée sur la **Figure 23B**. La courbe relativement stagnante montre que la position de l'image lors de l'entraînement n'influence pas la probabilité de commettre une erreur pour cette image.

2.2. Effet de primauté

Un résultat tout à fait intrigant concerne l'effet de primauté, observé sur le modèle FaRe uniquement pour des petites tailles de *datasets*. En effet, les graphiques obtenus avec $N = 30$ et $N = 20$ montrent que le modèle ne commet aucune erreur pour le premier item appris lors de l'entraînement. De plus, cette probabilité d'erreurs est inférieure à celle du dernier item étudié, suggérant un effet de primauté plus prononcé que l'effet de récence lorsque peu d'images sont présentées au modèle. Ces résultats rejoignent partiellement l'hypothèse de l'implication de la mémoire déclarative dans les processus de la MdR, et plus précisément de la familiarité. En effet, si les premiers items de la liste sont bel et bien encodés dans la mémoire à long terme comme mentionné précédemment, il n'est pas étonnant d'observer un effet de primauté lors d'une tâche de reconnaissance. Par ailleurs, Whittlesea (1993) a montré qu'un jugement de familiarité peut se baser sur un effet de primauté. Toutefois, l'étude a également montré que cet effet est moins prononcé que celui de récence.

Etonnement, l'effet de primauté disparaît lorsque l'on procède à une tâche de reconnaissance avec de plus grands *datasets* ; il est de même totalement absent des simulations avec le modèle Hebbien. Dès lors, les résultats laissent penser qu'un effet de primauté n'interviendrait pour faciliter un jugement de reconnaissance que sous certaines conditions. Pareillement, une étude sur des singes rhésus a montré la présence d'un effet de primauté pour des petits ensembles de données à apprendre et sa disparition lorsque le nombre de stimuli

augmente (Basile & Hampton, 2010). Selon les auteurs, ces résultats émaneraient d'un changement dans la stratégie de récupération, en passant de la familiarité pour les grands *datasets* à la recollection pour les petits *datasets*. Cette hypothèse est cohérente avec le fait que les sujets humains ont tendance à adopter naturellement une stratégie de récupération visant d'abord à la recollection des items aux extrémités de la série (Baddeley & Hitch, 1993).

Une façon de tester cette hypothèse chez l'homme serait de réaliser un test de reconnaissance avec un paradigme R/K/G sur deux tailles de *datasets*, une petite et une grande. Cela permettrait ainsi d'obtenir la proportion d'images recollectées et d'images reconnues sur base d'un SdF. Nous observerions ensuite la probabilité d'erreurs pour chaque item, à l'instar de ce qui a été réalisé lors de la Simulation 2 de ce mémoire. Nous supposons qu'un effet de primauté sera observé lorsque le nombre de stimuli est peu élevé, avec un taux de réponses R supérieur au taux de réponses K/G. Avec le *dataset* de grande taille, nous devrions constater un taux de réponse K/G plus élevé, en l'absence d'un effet de primauté.

3. Similarité visuelle

Commençons par comparer les performances du modèle FaRe avec celles du modèle Hebbien. Nous avons émis l'hypothèse que les performances chuteraient lorsque les images apprises par le modèle font partie de la même catégorie (dans le cas de la troisième simulation les chats). Plus spécifiquement, nous nous attendions à ce que cet effet soit plus marqué avec une règle d'apprentissage Hebbienne. Nos simulations semblent confirmer nos hypothèses en démontrant une différence significative entre les moyennes des scores entre des images similaires et dissimilaires, et ce pour les deux modèles. Par ailleurs, une analyse qualitative des données nous montre bel et bien que les performances du modèle FaRe sont très peu impactées par cet effet, contrairement à son homologue.

Cette différence entre les performances des modèles pourrait trouver une explication dans leurs caractéristiques intrinsèques. Comme expliqué, l'apprentissage anti-Hebbien fonctionnerait via l'extraction et l'apprentissage des caractéristiques individuelles propre à une image (Bogacz & Brown, 2003a). Pour illustrer ce mécanisme, prenons deux chats de races et de couleurs différentes (**Figure 29**). Le chat de gauche est blanc, a des yeux gris avec des pupilles ovales, ses deux pattes avant sont visibles sur la photo. Le chat de droite est noir, a des yeux jaunes avec des pupilles en fente ; il possède également une tache orange sur le museau. Dans le modèle anti-Hebbien, ce sont ces caractéristiques individuelles qui seront apprises pour chacun des chats ; elles seront ensuite utilisées par le modèle pour déterminer si oui ou non

l'image est familière. Le modèle Hebbien, quant à lui, apprendrait les caractéristiques générales d'une image, qui sont donc partagées avec une seconde image similaire (Bogacz et al., 2001b) : les deux chats ont des oreilles pointues, des yeux, un museau rose, etc.



Figure 29 | Deux images de chats reprises du « Cat Dataset » (W. Zhang et al., 2008).

Dans leurs travaux de modélisation de la catégorisation perceptuelle durant la petite enfance, French et coll. (French et al., 2001 ; Mareschal et al., 2000) ont exploré la présence d'une exclusivité asymétrique lors de la catégorisation des chiens et des chats. En résumé, un chien ne sera pas catégorisé comme un chat mais un chat pourrait être catégorisé comme un chien. Selon eux, cette asymétrie serait expliquée par le fait que les chiens présenteraient une plus grande distribution, c'est-à-dire une plus grande variété de caractéristiques perceptuelles, que les chats. Dans cette logique, si nous reproduisons la Simulation 3 uniquement avec des images de chiens, les performances du modèle Hebbien devraient être meilleures qu'avec les chats, étant donné que les premiers partageraient moins de caractéristiques communes. Les résultats de nos simulations ultérieurs semblent confirmer cette hypothèse et le modèle Hebbien discrimine mieux les chiens que les chats.

3.1. Erreur de familiarité et phénomène de Déjà-Vu

Revenons-en aux performances brutes du modèle Hebbien. Les résultats obtenus sont en faveur de l'hypothèse d'un effet de similarité visuelle sur la reconnaissance basée sur la familiarité, c'est-à-dire qu'une forte similarité entre des stimuli pourrait être à l'origine d'une illusion de familiarité (Cleary, 2008 ; Cleary et al., 2009). Dans le cadre de cette discussion, il est intéressant d'attirer l'attention sur un phénomène souvent délaissé par la communauté scientifique : l'expérience de Déjà-Vu (DV).

L'expérience de DV correspond à la sensation étrange de s'être déjà rendu dans un endroit ou d'avoir déjà fait une action, malgré la certitude du contraire (A.S. Brown, 2003). Selon certaines théories, le DV s'expliquerait par une forte impression de familiarité pour une situation présente, bien que cette situation n'ait jamais été vécue. En d'autres termes, cela pourrait correspondre à une illusion de familiarité éprouvée lors d'une reconnaissance, durant laquelle le sujet n'est pas capable d'identifier la source de cette reconnaissance (Cleary, 2008). Selon la théorie explicative du *Gestalt*, notre perception de la configuration d'une situation présente, qui serait similaire à celle d'une situation vécue précédemment, donnerait naissance à ce SdF. Le DV serait donc la conséquence de cette similarité entre une scène vécue précédemment et une scène vécue actuellement (Cleary et al., 2009). En faveur de cette hypothèse, des auteurs (Cleary et al., 2009) ont montré que quand une nouvelle scène ressemble à une ancienne d'un point de vue de la configuration des éléments qui la constituent, une augmentation du SdF vis-à-vis de la nouvelle scène se développe. Cela se produirait également lorsque l'ancienne scène ne parvient pas à être recollectée ; la probabilité d'apparition d'un DV semble alors accrue. En outre, le SdF serait plus intense s'il est accompagné d'un état de DV.

Des travaux antérieurs, basés sur des réseaux de neurones de Hopfield, ont mis en évidence un phénomène de DV en voulant modéliser le SdF (Bogacz et al., 2001a). Dans ce type de réseau, un stimulus mémorisé est récupérable via la stabilisation du réseau lorsqu'une partie adéquate de la cible stimule le réseau. On parle ainsi de réseau avec une mémoire adressable par son contenu (Hopfield, 1982). De plus, un stimulus possède également un état attracteur. Cet état permet au réseau de récupérer le stimulus même si ce dernier est légèrement différent. Dans leur modèle, la familiarité se modélise avant la récupération totale du stimulus, grâce à la fonction d'énergie qui devrait être plus faible pour les stimuli encodés que pour les nouveaux (Hertz et al., 1991). Bogacz et al. (2001a) ont constaté que si l'on présente au modèle des stimuli à l'état attracteurs, c'est-à-dire qu'ils ont été légèrement modifiés par rapport à ceux appris par le modèle, ces derniers peuvent engendrer des erreurs de familiarité. Ce type d'erreurs correspondrait à un DV car, par définition, le stimulus présente une forte familiarité alors qu'il n'a jamais été appris (A.S. Brown, 2003).

Dans notre modèle, une manière d'explorer ce phénomène serait donc de recenser les erreurs selon deux critères. Premièrement, les images familières qui ne sont pas reconnues ; deuxièmement, les nouvelles images qui sont considérées comme étant familière. Ce sont ces dernières qui correspondraient à un phénomène de DV (Bogacz et al., 2001a ; A.S. Brown, 2003). Ce type d'erreurs devrait par ailleurs être moins fréquent que les erreurs liées à la non-

reconnaissance d'anciennes images (Bogacz et al., 2001a). Pour apporter du soutien à l'hypothèse de la similarité comme cause du DV, la proportion de nouvelles images reconnues à tort comme étant familières devrait augmenter lorsque le modèle est testé avec des images similaires (Cleary et al., 2009). On pourrait également s'attendre à ce que la proportion d'erreurs du premier type – les anciennes images qui ne sont pas reconnues – reste stable.

3.2. Format du test

Un résultat tout à fait intrigant concerne les performances du modèle FaRe lorsqu'il est testé sur des images de même catégorie. En effet, si les analyses statistiques montrent une différence significative entre les deux groupes ($p < .001$), les performances du modèle FaRe ne s'effondrent pas pour autant. Ce dernier conserve une moyenne proche de 0, avec une précision de 99,6% pour 40 images apprises. Par ailleurs, les résultats montrent que la taille d'effet pour le modèle FaRe ($r = -0.7$) est inférieure à celle obtenue pour le modèle Hebbien ($r = -1$). D'un point de vue qualitatif, ces résultats vont à l'encontre de notre hypothèse concernant l'impact négatif de la similarité des stimuli sur les performances du modèle FaRe.

Des auteurs (Migo et al., 2009) ont montré que cet effet de la similarité dépend essentiellement du format du test utilisé. Plus précisément, ils ont observé, chez un patient atteint de lésions hippocampiques avec des mécanismes de recollection déficitaires, des performances préservées lors d'une tâche de RCF où les leurres sont fortement similaires aux cibles. A contrario, les performances de ce même patient lors d'une tâche de reconnaissance Oui/Non sont significativement impactées par la similarité des leurres. Cela peut s'expliquer par le fait que, lors d'un test de RCF, une image familière produira tout de même un signal de familiarité plus puissant qu'une nouvelle image, malgré la forte similarité entre les deux (Norman, 2010). Dans une tâche de reconnaissance Oui/Non, le patient ne peut pas s'appuyer sur cette différence dans le degré de familiarité étant donné que les images sont présentées une par une. Ainsi, la recollection serait nécessaire à la bonne réalisation d'une tâche de reconnaissance Oui/Non où les leurres et les cibles sont similaires (Westerberg et al., 2006).

En conclusion, un test de RCF ne serait pas approprié pour mettre en évidence un effet de la similarité sur les performances. En revanche, les performances lors d'une tâche de reconnaissance Oui/Non devraient donc être plus sensibles à cette similarité inter-items. Nous avons testé l'hypothèse d'un potentiel effet du format du test en implémentant un test de reconnaissance Oui/Non, que nous avons administré au modèle FaRe. Notre prédiction était que les performances du modèle seraient réduites lorsque la tâche est administrée avec des

images de catégorie identique comparées à des images de catégories différentes. Les détails méthodologiques de cette simulation sont décrits dans l'**Annexe 3**. Nos simulations préliminaires montrent que le modèle est moins performant lors d'une tâche de reconnaissance Oui/Non et ce, dans les deux conditions de similarité (**Figure A5**). La différence de performances pour les images similaires et dissimilaires lors de la tâche de reconnaissance Oui/Non semble correspondre à celle obtenue pendant la tâche de RCF. Actuellement, le modèle FaRe échoue donc à montrer un effet du format du test sur l'hypothèse de la similarité, ce qui contredit les données de la littérature (Migo et al., 2009 ; Westerberg et al., 2006). Toutefois, étant donné l'imperfection de cette modélisation préliminaire, de plus amples simulations mériteraient d'être réalisées pour approfondir ce point.

4. Familiarité des visages humains

Les résultats de la dernière simulation semblent indiquer que le modèle est capable de correctement discriminer les visages connus des visages inconnus, malgré le fait que les visages familiers soient présentés sous un jour différent. De surcroît, ces résultats sont obtenus sans avoir eu recours à des synapses complexes (Ji-An et al., 2022). Nos précisions d'environ 80% sont néanmoins inférieures à celles obtenues par de tels modèles.

Plusieurs études, tant chez l'homme et l'animal, ont déjà identifié le cortex périrhinal comme une zone clé dans la détection de visages familiers (Martin et al., 2013 ; She et al., 2021). D'autres zones cérébrales peuvent aussi jouer un rôle dans cette familiarité pour les visages, tel que cortex temporal inférieur par exemple (Tsao et al., 2006). Chez les souris, l'aire CA2 de l'hippocampe semble également impliquée dans la détection de la familiarité (Boyle et al., 2022). Enfin, d'autres études ont montré l'impact positif de la composante émotionnelle sur les jugements de familiarité (Duke et al., 2014 ; Fiacconi et al., 2016). Prises ensemble, ces données suggèrent donc que le cortex périrhinal n'est pas la seule structure cérébrale à intervenir lorsqu'on est confronté à un jugement de familiarité pour d'autres personnes.

Les modèles testés dans ce mémoire n'implémentent que la contribution du cortex périrhinal au SdF. Ils excluent dès lors toute autre contribution des régions qui projettent sur le cortex périrhinal lors de la reconnaissance en temps réel. Il n'est donc pas étonnant que leur précision ne soit pas aussi élevée que celle du modèle de Ji-An et al. (2022). Cela renforce le constat de la complexité des mécanismes de familiarité et la nécessité de modèles plus élaborés pour rendre compte de ce phénomène de façon plus complète.

Nous relevons néanmoins deux limites liées à cette dernière simulation, effectuée avec des photos de célébrités. Y remédier pourrait contribuer à améliorer les performances des modèles.

4.1. Pré-entraînement du réseau

Premièrement, afin d'extraire les caractéristiques d'un visage, Ji-An et al. (2022) utilisent un RPC spécialement pré-entraîné pour la détection de visages avec le *dataset* « Ms-celeb-1m » (Guo et al., 2016). De plus, Kazanovich et Borisyuk (2021) ont fait l'hypothèse que la réussite lors d'une tâche de reconnaissance dépend notamment du pré-entraînement subi par le modèle. Selon eux, si le système d'extraction de caractéristiques est optimisé par un entraînement préliminaire, un réseau de neurones, aussi basique soit-il, devrait être capable de reconnaître sans difficulté un grand nombre de stimuli. Dans notre modèle tout comme dans le modèle FaRe, le RPC a bel et bien été entraîné via ImageNet mais pas spécifiquement avec des visages. Il pourrait donc être intéressant de reproduire la simulation 4 avec un RPC spécialement conçu, ou en tout cas pré-entraîné, pour la reconnaissance faciale.

4.2. Dataset chaotique

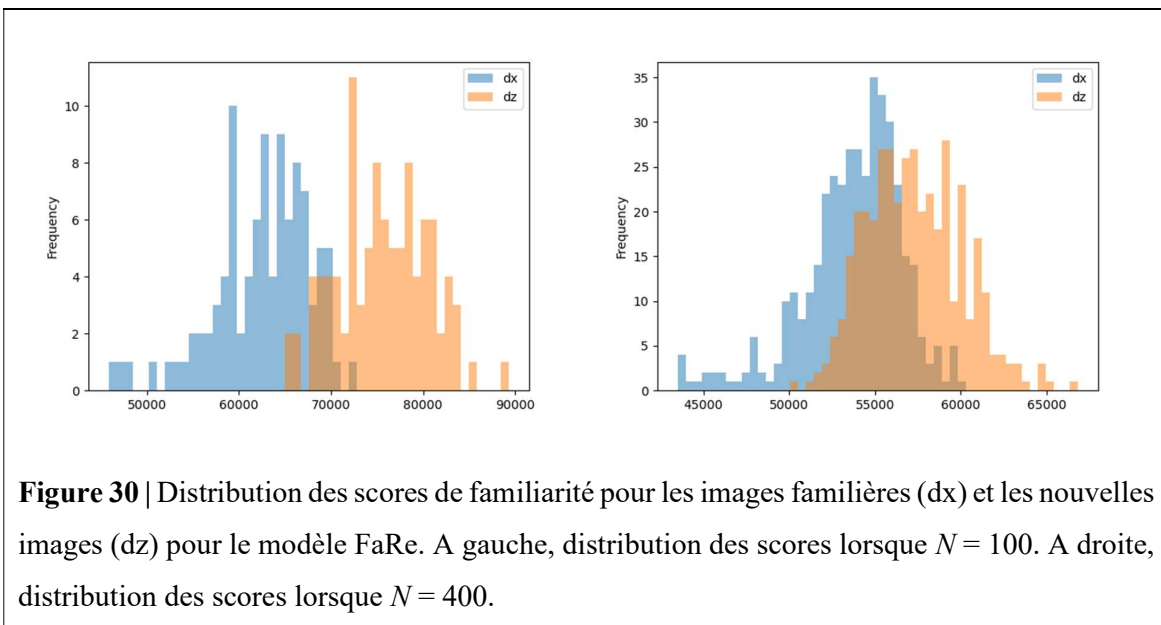
Une autre limite de notre simulation concerne le *dataset* utilisé en lui-même (Parkhi et al., 2015). Les images étant soumises à des droits d'auteurs, ils n'ont pas pu les rendre directement accessibles dans un fichier. A la place, Parkhi et al. (2015) fournissent un document texte contenant l'URL associée à chacune des images. Malheureusement, beaucoup d'URL n'existent plus et renvoient dès lors à des images automatiques. Ensuite, une fois téléchargé, le *dataset* est constitué de plusieurs dossiers, chacun correspond à une célébrité. Etonnement, des photos de personnes différentes ont été insérées automatiquement dans le dossier d'une célébrité en particulier. C'est notamment le cas d'acteurs, qui retrouvent des photos de leurs co-stars dans leur dossier. Ainsi, lors de la phase de test, il est probable que le modèle ait sélectionné la photo de l'un de ces intrus à la place d'une image différente de la célébrité sur laquelle le modèle s'est entraîné. Finalement, un grand nombre d'images correspond en réalité à des photos groupées, sur lesquelles se trouve un grand nombre de personnes. Si une telle photo a été sélectionnée lors de l'entraînement, il est difficile de déterminer avec certitude si les traits de la personne cible ont été correctement appris par le modèle ou si ce n'est pas plutôt la photo dans son ensemble qui a été encodée.

5. Comparaison entre les modèles

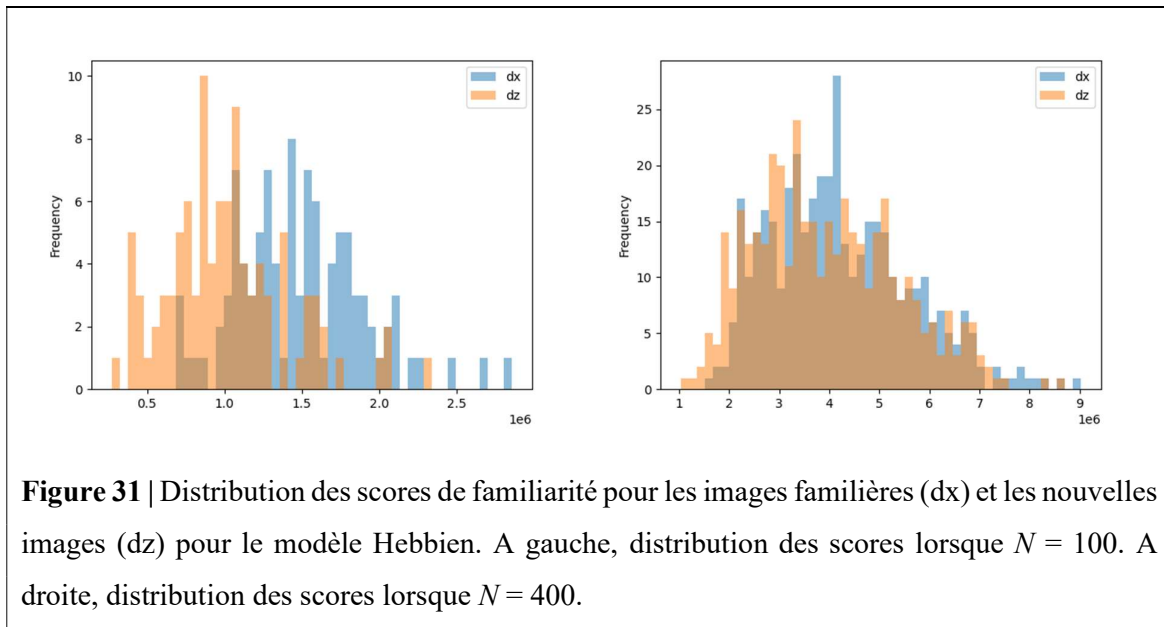
Au vu de la différence des résultats obtenus dans les simulations, il est maintenant évident que le modèle FaRe se comporte différemment du modèle Hebbien. Cela ne signifie pas pour autant qu'un des modèles soit plus exacte que l'autre dans sa modélisation de la familiarité. En comparant ci-dessous les deux modèles, nous espérons au contraire défendre leur complémentarité pour décrire correctement un phénomène aussi complexe que le SdF.

5.1. Nature de la distribution

La première comparaison repose sur la représentation graphique des distributions des scores de familiarité pour les cibles (dx) ainsi que pour les leurres (dz). L'analyse de la forme des courbes obtenues avec le modèle FaRe (**Figure 30**) nous montre deux courbes Gaussiennes, qui se chevauchent partiellement. Cette observation est cohérente avec le modèle DPSD de Yonelinas et coll. (Yonelinas et al., 1996), qui décrit la familiarité comme un processus de détection de signaux, où les leurres et les cibles prennent la forme de deux courbes en cloche qui s'entrecroisent (voir **Figure 2**). Dans la partie droite du graphique, on remarque également que lorsque le nombre d'images apprises par le modèle augmente, les deux courbes se chevauchent de plus en plus. Cela concorde avec le nombre d'erreurs de reconnaissance croissant pour les plus grandes tailles de *datasets*. Pour apporter un soutien supplémentaire au modèle FaRe, nous pourrions calculer les courbes ROC correspondantes. Si ces dernières coïncident avec celles obtenues dans la littérature (voir **Figure 3**), cela conforte encore une fois l'hypothèse des deux processus distincts de recollection et de familiarité.



En analysant les courbes de la **Figure 31**, nous pouvons constater que lorsque $N = 100$, les scores forment des courbes qui ressemblent plus ou moins à une distribution standard. Ceci pourrait suggérer que la familiarité s'exprime comme un processus de détection de signaux dans le modèle Hebbien. Cependant, lorsqu'on augmente le nombre de stimuli appris par le modèle, on constate que les courbes se chevauchent presque entièrement ; une légère asymétrie vers la droite (i.e. *positive skew*) est par ailleurs détectable.



L'asymétrie droite pourrait s'expliquer par l'explosion des poids durant l'apprentissage. En effet, ces derniers prennent rapidement des valeurs exponentielles au fur et à mesure que le nombre d'images augmente. Il n'est dès lors pas impossible que ces poids aux valeurs parfois aberrantes produisent des sorties qui s'éloignent progressivement de la moyenne, créant ainsi cette asymétrie vers la droite. Cette distribution est par ailleurs cohérente avec la distribution des potentiels membranaires en sortie du modèle Hebbien qui montre une dissymétrie vers la gauche (voir **Figure 15**).

Concernant le chevauchement entre les couches, il n'est pas étonnant d'observer une distribution de la sorte au vu du taux d'erreurs observé avec le modèle Hebbien lorsque la taille du *dataset* excède un certain nombre. En effet, comme expliqué, plus le recouvrement est grand et plus le modèle aurait du mal à déterminer si une image est ancienne ou nouvelle. Cet effet semble inhérent au fonctionnement du modèle Hebbien. Si ce dernier apprend la structure commune aux différents stimuli qui lui sont présentés, plus le nombre d'items est élevé et plus il aura tendance à généraliser cette structure à tous les stimuli. En d'autres termes, quand le

nombre d'images dépasse un certain seuil, le modèle Hebbien apprend uniquement une structure partagée par toutes les images du *training set*, ce qui l'empêcherait de discriminer correctement.

Etrangement, un effet similaire a été constaté par Norman & O'Reilly (2003) avec leur modèle hippocampique qui, pour rappel, avait pour but la modélisation de la recollection. Dans leur modèle, si on augmente le recouvrement entre les items, les courbes prennent progressivement une forme Gaussienne⁶ et leur chevauchement augmente. Ils expliquent ce phénomène par l'idée que le modèle hippocampique ne parviendrait plus à séparer correctement les *patterns* qui composent les entrées ; l'hippocampe perdrait sa capacité à assigner des représentations distinctes aux *patterns* d'entrées qui lui sont fournis. Cela rejoint notre hypothèse selon laquelle le modèle Hebbien n'apprendrait que des représentations prototypiques des stimuli lorsqu'un trop grand nombre d'images lui sont présentées, compliquant ainsi la prise de décision lors de la tâche de reconnaissance.

5.2. Plasticité synaptique

La principale différence entre les deux modèles présentés dans ce travail concerne bien entendu les mécanismes de plasticité synaptique qui ont été modélisés. Rappelons que dans les deux cas, ces mécanismes ont été simplifiés pour concorder à des neurones artificiels simples, comme ceux décrits par McCulloch & Pitts (1943). Par ailleurs, dans notre modèle Hebbien, nous avons considéré les neurones d'entrée ayant une valeur positive comme étant actifs et ceux avec une valeur négative comme étant au repos. Le **Tableau 2** reprend les différents mécanismes de plasticité simulés par les modèles ainsi que le sens de la modification des poids (augmentation ou diminution).

Considérons d'abord le modèle FaRe. Les synapses du modèle sont déprimées lorsqu'une image est mémorisée. Pour rappel, cela correspond au mécanisme de LTD homosynaptique, mécanisme qui a bel et bien été démontré dans le cortex périrhinal (Cho et al., 2000). Par ailleurs, ce type de plasticité est plus simple à reproduire dans des tranches de cortex périrhinal si le neurone post-synaptique est dépolarisé que s'il est hyperpolarisé (Cho et al., 2000). Cette constatation est cohérente avec le module anti-Hebbien étant donné la modification des poids des neurones actifs et l'absence de modification pour les neurones inactifs. Dans ce cas de figure, l'activité de la couche de neurones de nouveauté est plus faible pour les stimuli

⁶ Dans le modèle hippocampique, la distribution des scores de familiarité ne se présente pas sous la forme d'une courbe en cloche (voir la Figure 6 dans Norman & O'Reilly (2003), p. 620).

familiers, ce qui correspond au phénomène de répétition par suppression observé dans la littérature (Grill-Spector et al., 2006 ; Meyer & Rust, 2018). En cela, le module de mémoire anti-Hebbien correspond aux observations expérimentales dans le cortex périrhinal.

Pour maintenir l'excitabilité neuronale, le module anti-Hebbien va également devoir potentialiser les neurones d'entrée inactifs – ceux ayant une valeur négative – via un mécanisme de LPT hétéro-synaptique. Dans le cerveau, ce mécanisme de potentialisation est homo-synaptique et nécessite l'activité des deux neurones pré- et post-synaptique, ce qui ne correspond pas aux observations du modèle anti-Hebbien. Une explication biologique à cette activité synaptique atypique pourrait résider dans la désinhibition : la plasticité du modèle anti-Hebbien pourrait ainsi être interprétée comme une potentialisation des synapses inhibitrices telle que décrite par Schulz et al. (2021). La plasticité inhibitrice va augmenter l'inhibition des neurones excitateurs correspondant à un stimulus ; cette inhibition restera faible pour les nouveaux stimuli en comparaison aux stimuli familiers (Schulz et al., 2021).

Passons maintenant au modèle Hebbien. Dans cet algorithme, les synapses inhibitrices sont potentialisées en réponse à un stimulus, reproduisant ainsi le mécanisme de LPT homo-synaptique. De plus, les poids entre les entrées actives et les neurones de nouveauté inactifs sont également déprimés, mimant ainsi le mécanisme de LTD homo-synaptique. Dans la situation inverse – entrées inactives mais neurones de nouveauté actifs – ce sont les mécanismes de LTD hétéro-synaptiques qui sont reproduits. Bien que ces derniers n'aient pas été observés dans le cortex périrhinal, ils ont néanmoins été découverts dans d'autres parties du cerveau (Ito, 1989). Théoriquement, cette reproduction plus complète de la plasticité synaptique confère au modèle Hebbien une plausibilité biologique supérieure à son homologue anti-Hebbien.

Cependant, certaines variables inhérentes à la plasticité synaptique Hebbienne dans le cerveau n'ont pas été prises en compte. En effet, dans le cerveau humain, des paramètres temporels très précis sont essentiels au déclenchement des potentiel d'action des neurones (L. I. Zhang et al., 1998). A titre d'exemple, L.I. Zhang et al. (1998) ont montré que quand un signal provenant du neurone pré-synaptique arrive 20ms avant ou pendant le pic du potentiel d'action du neurone post-synaptique, la synapse est potentialisée par LTP. A l'opposé, quand le signal pré-synaptique arrive 20ms après le pic du neurone post-synaptique, la synapse est déprimée par LTD (**Figure 32**).

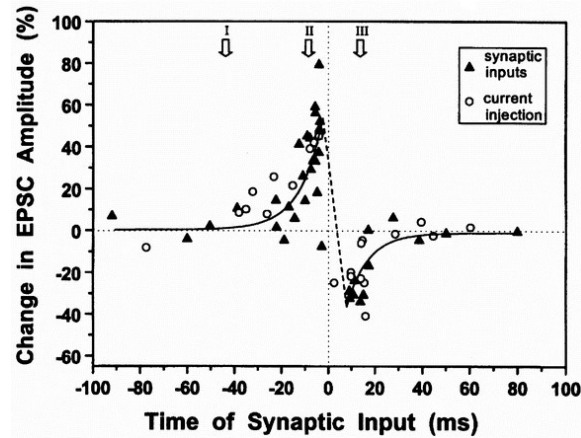


Figure 32 | Fenêtres critiques pour la potentialisation et la dépression (L. I. Zhang et al., 1998). Le graphique représente le pourcentage de variations de l'amplitude du courant excitateur post-synaptique évoqué sur le déclenchement du potentiel d'action post-synaptique par rapport au pic d'un potentiel d'action.

Ainsi, un algorithme Hebbien plus complet devrait prendre en compte des paramètres tels que le temps de déclenchement du neurone pré-synaptique, le délai de transmission entre ce déclenchement et son effet sur le neurone post-synaptique ou encore le temps de déclenchement du neurone post-synaptique, comme fonction de la modification des poids w_{ij} (Sougné & French, 2001), cela afin de reproduire les relations temporelles inhérentes à la plasticité synaptique.

Finalement, le modèle Hebbien tel que conçu par Bogacz et al. (2001b) implique la participation d'une troisième couche de neurones, correspondant à la projection des interneurons inhibiteurs sur les neurones de nouveauté (voir Partie 1, Section 6.4). Lors de l'apprentissage, les interneurons vont inhiber l'activité initiale d'une image lorsqu'elle est présentée pour la première fois. Cette troisième couche permet donc la reproduction de la diminution de l'activité des neurones dans le cortex périrhinal. Elle n'a toutefois pas été modélisée dans le cadre de ce travail. En effet, par simplicité, nous avons pris en considération l'activité des interneurons inhibiteurs comme évaluateurs de la familiarité, alternative suggérée par Bogacz & Brown (2003b). De ce fait, il est approprié de considérer notre modèle Hebbien comme incomplet. Des simulations ultérieures, tenant compte de cette troisième couche, devraient être réalisées afin de vérifier si les performances parfois décevantes de notre modèle ne découlent pas de la méthode de modélisation choisie.

Tableau 2 | Mécanismes de plasticité synaptique reproduits par les modèles.

Modèle	Entrée	Sortie	Sens de la modification	Mécanisme
Hebbien ($\eta > 0$)	+	+	↑	LTP
	+	-	↓	LTD homo
	-	+	↓	LTD hétéro
	-	-	↑	LTP
anti-Hebbien ($\eta < 0$)	+	+	↓	LTD
	+	-	/	/
	-	+	↑	LTP hétéro
	-	-	/	/

Note. Le signe + signifie que le neurone est actif ; le signe moins signifie que le neurone est inactif. η correspond à la constante d'apprentissage.

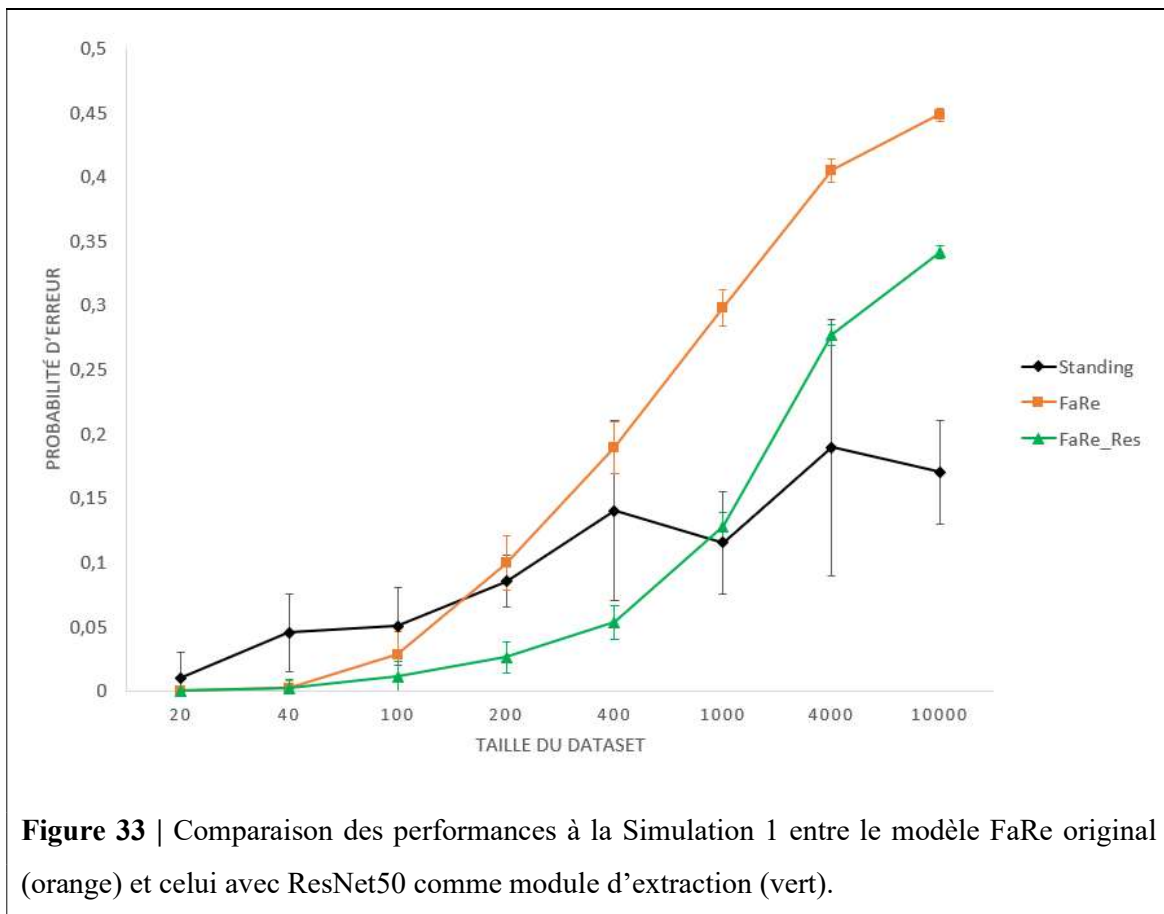
En conclusion, si le modèle anti-Hebbien semble souffrir de certaines lacunes pour affirmer sa plausibilité biologique, il semble toutefois plus prometteur que son homologue Hebbien – tel que modélisé dans ce mémoire – pour expliquer certaines données comportementales. Récemment, un modèle utilisant un algorithme de méta-apprentissage (i.e. *meta-learning*) de plasticité synaptique pour modéliser la familiarité convergeait plus fréquemment vers cette règle d'apprentissage anti-Hebbienne (Tyulmankov et al., 2022). Les auteurs en ont conclu que la plasticité synaptique découlant du modèle anti-Hebbien est un mécanisme biologiquement plausible pour l'apprentissage de stimuli. Toutefois, avant de conclure définitivement à la supériorité du modèle anti-Hebbien afin de modéliser le SdF, de plus amples simulations mériteraient d'être réalisées avec un modèle Hebbien plus complet.

5.3. Module d'extraction

Clôturons cette comparaison avec un des résultats les plus intéressants de la Simulation 1, à savoir l'amélioration des performances du modèle Hebbien lorsque le module d'extraction utilisé est plus performant. Kazanovich & Borisyuk (2021) ont déjà insisté sur l'importance cruciale du module d'extraction de caractéristiques pour un jugement de familiarité adéquat. Selon eux, une décision correcte lors d'une tâche de reconnaissance dépend de l'efficacité du système d'extraction de caractéristiques ainsi que du pré-entraînement que ce système a subi. Chez l'homme, cela correspondrait au vécu durant les stades précoces du développement du cerveau lors de la petite enfance ; ces expériences façonneraient le système visuel de façon à

pouvoir encoder les images du quotidien de façon optimale. Les résultats de nos simulations avec le modèle Hebbien appuient cette hypothèse et les performances du réseau convolutif utilisé permettent une amélioration des performances du module de mémoire.

Pour aller plus loin, nous avons reproduit la Simulation 1 avec le modèle anti-Hebbien en utilisant non plus AlexNet mais bien ResNet50 comme module d'extraction. Les résultats sont observables sur la **Figure 33**, mis en rapport avec ceux de la simulation initiale du modèle FaRe. Comme attendu, les performances du modèle sont améliorées de façon significative. En particulier, la courbe se rapproche fortement des données expérimentales lorsque $N = 1000$ et correspond à l'écart-type supérieur chez l'homme lorsque $N = 4000$ et l'écart-type inférieur lorsque $N = 400$. Le modèle garde cette tendance à sous-estimer les performances lorsque $N = 10000$ mais les scores se rapprochent tout de même de la courbe expérimentale. Pour les *datasets* de plus petites tailles, le modèle montre une surestimation des performances par rapport aux humains.



Pour approfondir l’exploration du module d’extraction, nous pourrions analyser les performances du modèle avec des *patterns* d’images abstraits. Des preuves expérimentales montrent effectivement que la MdR visuelle, et plus spécifiquement la familiarité, est plus fiable et rapide avec des images naturelles porteuses de significations (Bellhouse-King & Standing, 2007). Il a été démontré que la performance des sujets lors d’une tâche de reconnaissance est bien meilleure pour les images naturelles que pour les images abstraites et/ou artificielles, lesquelles sont dénuées de sens. Dans les simulations de Kazanovich & Borisjuk (2021) explorant le sujet, le modèle FaRe s’est montré moins performant lorsque testé sur des images constituées d’un ensemble de figures géométriques. Selon les auteurs, notre système visuel, qui se développe au fur et à mesure de l’enfance, est inefficace pour coder ces images artificielles, dénuées de significations. Pour apporter du soutien à nos hypothèses, nous devrions également tester le modèle FaRe avec ResNet50 comme module d’extraction ainsi que le modèle Hebbien sur ces *patterns* abstraits. Si les performances sont significativement inférieures à celles obtenues avec des images naturelles, cela appuierait notre hypothèse.

6. Limites et implications

Avant de conclure, nous épinglerons certaines limites inhérentes non seulement à mon implémentation et plus généralement au domaine de l’IA.

6.1. Simplicité du modèle

Pourquoi faire compliqué quand on peut faire simple ? Ce principe du Rasoir d’Ockham pourrait être appliqué à notre modèle, qui parvient à reproduire adéquatement des données expérimentales via un simple réseau de neurones, et qui apprend les caractéristiques d’une image grâce à une règle d’apprentissage aussi basique que celle de Hebb (1949). C’est la simplicité du modèle FaRe – ainsi que de notre variante Hebbien – qui lui donne son élégance. Toutefois, aussi élégants soient-ils, ces modèles restent incomplets. Parfois, la simplicité n’est pas conforme à la nature, comme nous le démontre la complexité du cerveau par exemple.

La familiarité – et plus généralement la reconnaissance – est un phénomène complexe ; elle implique la participation de plusieurs aires cérébrales et requiert dès lors des modèles théoriques de plus en plus complexes (Bastin et al., 2019). Nos données vont dans le même sens car elles échouent à reproduire avec exactitude les courbes obtenues dans la littérature, notamment avec des grandes tailles de *training set* (voir **Figures 20 et 21**). Selon nous, les modèles proposés resteront incomplets tant qu’ils ne tiendront pas compte d’autres structures et mécanismes, tels que l’hippocampe et la recollection, impliqués lors de la reconnaissance

(Norman & O'Reilly, 2003). Par ailleurs, une implémentation plus réaliste des réseaux neuronaux, en utilisant des neurones continus, permettrait d'outrepasser certaines faiblesses du modèle Hebbien tel que l'oubli catastrophique (Sougné & French, 2001).

6.2. Apprentissage du modèle

La question légitime à poser pour tous types de réseaux pourrait être : qu'est-ce qui est réellement appris par le modèle ? Pour illustrer notre propos, prenons cet exemple tiré du livre *Artificial Intelligence : A guide for Thinking Humans* de Mélanie Mitchell (2019). Elle nous présente un réseau de neurones permettant de différencier deux catégories d'images, à savoir « contient un animal » et « ne contient pas un animal » (**Figure 34**).



Figure 34 | Illustrations d'un animal et d'un paysage (catégorie « non-animal ») présentées lors d'une tâche de classification (Mitchell, 2019).

Étonnement, dans certaines situations, le réseau se trompe d'une façon absolument absurde pour un être humain : lorsque l'image présente un arrière-plan flouté, le réseau classe l'image dans la catégorie « contient un animal » même lorsqu'aucun animal n'est présent sur la photo. En effet, quand on photographie un animal, on fait la mise au point sur l'animal en question et on obtient dès lors un arrière-plan flouté. Alors que pour un paysage, par exemple, le photographe essaye plutôt d'avoir une grande profondeur de champ, avec un arrière-fond aussi net que possible. En réalité, le réseau présenté ici a uniquement appris à détecter si l'horizon est flou ou net. En résumé, si on photographie un animal en faisant la mise au point sur l'arrière-plan plutôt que sur la bête, le réseau catégorise la photo comme ne contenant pas d'animal.

Il s'agit là d'une limite prépondérante aux réseaux convolutifs tels qu'utilisés dans ce mémoire : « la machine apprend ce qu'elle observe dans les données plutôt que ce que nous, humain, pourrions observer » (Mitchell, 2019, p105). A supposer qu'il y ait des régularités statistiques dans un ensemble de données, le réseau pourrait apprendre ces régularités plutôt

que ce qu'il devrait réellement apprendre. Cette formulation semble parfaitement refléter le fonctionnement du module Hebbien, qui apprend les caractéristiques communes à plusieurs images. En conclusion, il est pertinent de penser que notre modèle Hebbien pour la familiarité n'apprend pas les images mais seulement une association statistique entre celles-ci. Cela expliquerait ainsi ses plus faibles performances comparées au modèle FaRe.

6.3. Code en open access

Un dernier point sur lequel il nous paraît important d'attirer l'attention concerne *l'open access*. Dans le cadre de ce travail, nous avons été amenés à répliquer le modèle informatique proposé par Kazanovich et Borisjuk (2021). Lors de cette implémentation, le code a été écrit en respectant scrupuleusement les indications reprises dans l'article. Néanmoins, les résultats obtenus ne permettaient pas la réplication exacte de leurs données. Contacté, le Dr. R. Borisjuk, que je remercie pour sa réponse, a confirmé que leur code n'est pas disponible en ligne. Afin de faire fonctionner le modèle, nous avons dû ajouter au code des mécanismes de compétition et d'inhibition (**Annexe 2**) qui n'étaient pas mentionnés dans leur article⁷. En définitive, les performances de notre implémentation finale du modèle FaRe se rapprochent de celles des auteurs et des données expérimentales (voir **Figure 22**), mais force est de constater qu'elles ne sont pas exactement similaires.

Cet exemple nous invite à défendre la réelle nécessité pour les chercheurs en IA de publier leur code en *open access*. La recherche scientifique passe majoritairement par la réplication des données ou des expériences. Si les chercheurs ne peuvent pas s'appuyer sur des bases complètes afin de répliquer les précédentes modélisations et ainsi pouvoir améliorer les modèles d'apprentissage machine, les avancées en seront ralenties. La science progressera plus rapidement si toutes les informations sont disponibles d'emblée. Par ailleurs, c'est en ayant accès à l'intérieur des modèles que l'on peut les protéger de certains biais sociétaux, directement transmis par l'être humain, et qui peuvent être lourds de conséquences pour certaines ethnies (voir Mitchell, 2019, p. 106-109).

⁷ Ces mécanismes étaient néanmoins décrits dans le papier d'Androulidakis et al. (2008), dont les codes n'étaient pas non plus disponibles en open access.

CONCLUSIONS

Ce travail est, à notre connaissance, le seul à avoir combiné réseaux convolutifs et apprentissage Hebbien pour modéliser la familiarité. Les résultats de nos simulations montrent que les performances de notre modèle sont moindres en comparaison au modèle FaRe (Kazanovich & Borisyuk, 2021). Ce dernier présente effectivement une meilleure capacité de stockage et semble plus résistant face à l'oubli catastrophique, alors même qu'il n'est pas pourvu de neurones continus. Le modèle FaRe parvient de surcroît à rendre compte de plusieurs phénomènes comportementaux tels que l'effet de récence – et de primauté sous certaines conditions – alors que notre modèle Hebbien peine à montrer un effet de primauté marqué par une courbe élégante (Basile & Hampton, 2010 ; Whittlesea, 1993). Finalement, à l'inverse du modèle FaRe, le modèle Hebbien est fortement impacté par la similarité entre les stimuli lors d'une tâche de RCF, ce qui va à l'encontre des données expérimentales (Migo et al., 2009 ; Westerberg et al., 2006). Nos simulations préliminaires sur ce dernier devraient être approfondies afin de confirmer l'absence d'un effet du format du test.

Doit-on pour autant renier le module de mémoire Hebbien et ses mécanismes de plasticité synaptique comme explication biologique de la familiarité ? Selon nous, la réponse à cette question est négative. Le modèle Hebbien fait preuve de bonnes performances lorsque le nombre d'images apprises en mémoire n'est pas trop élevé ; dans ce même cas de figure, il parvient en outre à reproduire les courbes Gaussiennes caractéristiques de la DPSD (Yonelinas, 1994). Comme mentionné à plusieurs reprises, l'apprentissage Hebbien se baserait sur les caractéristiques communes, partagées par plusieurs stimuli (Bogacz & Brown, 2003b). Chez l'homme, il n'est impossible que ce mode de fonctionnement pour effectuer un jugement de familiarité soit efficace lorsqu'il est confronté à un nombre limité d'informations. A l'inverse, face à trop d'informations, la familiarité pourrait s'exprimer au travers de détails plus caractéristiques d'une information particulière, comme dans le modèle FaRe et son apprentissage anti-Hebbien.

La familiarité est un phénomène complexe qui, comme le suggèrent certains auteurs (Duke et al., 2014 ; Fiacconi et al., 2016), peut émerger de différentes sources, qu'elles soient mnésiques, proprioceptives ou encore affectives. Si la modélisation nous aide à comprendre certains mécanismes du SdF, elle ne pourra pas apporter à elle seule rendre compte de tous les aspects de ce phénomène. Particulièrement, les modèles proposés dans ce travail, bien que prometteurs, sont trop simplistes et restent pour l'instant incomplets.

Au fur et à mesure de cette discussion, plusieurs pistes ont été suggérées afin d'améliorer notre modèle Hebbien ou encore d'explorer plus en profondeur certaines de nos hypothèses. Quelques-unes de ces suggestions sont rappelées ci-dessous :

- tester le modèle Hebbien sur des *patterns* abstraits et/ou des images artificielles, à l'image de ce qui a été réalisé avec le modèle FaRe par Kazanovich & Borisjuk (2021) ;
- proposer un nouveau module de mémoire Hebbien, plus complet, qui tient compte de la projection des interneurons inhibiteurs sur les neurones de nouveauté ; ce modèle devrait selon nous utiliser des neurones continus, qui prennent en compte les paramètres temporels pour la plasticité synaptique ;
- calculer les courbes ROC à partir de la distribution des scores de familiarité obtenue avec le modèle FaRe, afin d'apporter une crédibilité supplémentaire à ce dernier dans la modélisation du SdF ;
- améliorer la qualité méthodologique de la modélisation d'un test de reconnaissance Oui/Non en vue d'explorer plus en profondeur un potentiel effet du format du test.

Enfin, un aspect fondamental de la modélisation cognitive réside dans la capacité des modèles à réaliser des prédictions testables chez l'homme (Zuidema et al., 2020). Ici, le modèle FaRe prédit que le nombre d'items conditionne l'apparition de l'effet de primauté, possiblement en raison d'un changement dans la stratégie de récupération (Basile & Hampton, 2010). Pour tester cette prédiction chez l'homme, on pourrait réaliser un test de reconnaissance Oui/Non couplé à un paradigme R/K/G sur deux ensembles de données (petit et grand). L'analyse des probabilités d'erreurs et des taux de réponse R/K/G au fil de la tâche nous permettrait d'éprouver l'hypothèse d'un changement de stratégie de récupération selon le nombre de stimuli.

En conclusion, ce travail a le mérite d'explorer une grande variété de données liées à la familiarité, ouvrant ainsi la porte à de nouvelles hypothèses et expérimentations. Nous regrettons que le temps nous ait manqué pour en franchir la plupart. Dans notre philosophie scientifique, nous mettons cependant les codes à disposition des chercheurs désireux de poursuivre sur cette route.

ANNEXES
& BIBLIOGRAPHIE

BIBLIOGRAPHIE

- Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences*, 22(3), 425–444.
<https://doi.org/10.1017/S0140525X99002034>
- 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*, 123(4), 800–815.
<https://doi.org/10.1093/brain/123.4.800>
- Aggleton, J. P., & Saunders, R. C. (1997). The Relationships Between Temporal Lobe and Diencephalic Structures Implicated in Anterograde Amnesia. *Memory*, 5(2), 49–71.
<https://doi.org/10.1080/741941143>
- Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., & Yonelinas, A. P. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia*, 43(12), 1810–1823.
<https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2005.01.019>
- Androulidakis, Z., Lulham, A., Bogacz, R., & Brown, M. W. (2008). Computational models can replicate the capacity of human recognition memory. *Network: Computation in Neural Systems*, 19(3), 161–182. <https://doi.org/10.1080/09548980802412638>
- Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, 21(2), 146–155. <https://doi.org/10.3758/BF03202726>
- Basile, B. M., & Hampton, R. R. (2010). Rhesus monkeys (*Macaca mulatta*) show robust primacy and recency in memory for lists from small, but not large, image sets. *Behavioural Processes*, 83(2), 183–190. <https://doi.org/10.1016/j.beproc.2009.12.013>
- Bastin, C., Besson, G., Simon, J., Delhay, E., Geurten, M., Willems, S., & Salmon, E. (2019). An Integrative Memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X19000621>
- Bellhouse-King, M. W., & Standing, L. G. (2007). Recognition memory for concrete, regular abstract, and diverse abstract pictures. *Perceptual and Motor Skills*, 104(3), 758–762.
<https://doi.org/10.2466/1>
- Besson, G., Ceccaldi, M., & Barbeau, E. J. (2012). L'évaluation des processus de la mémoire de reconnaissance. *Rev Neuropsychol*, 4(4), 242–254. <https://doi.org/10.1684/nrp.2012.0238>

- Blalock, L. D. (2015). Stimulus familiarity improves consolidation of visual working memory representations. *Attention, Perception, and Psychophysics*, 77(4), 1143–1158.
<https://doi.org/10.3758/S13414-014-0823-Z>
- Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407), 31–39. <https://doi.org/10.1038/361031a0>
- Bogacz, R., & Brown, M. W. (2002). The restricted influence of sparseness of coding on the capacity of familiarity discrimination networks. *Network: Computation in Neural Systems*, 13(4), 457–485. https://doi.org/10.1088/0954-898X_13_4_303
- Bogacz, R., & Brown, M. W. (2003a). An anti-Hebbian model of familiarity discrimination in the perirhinal cortex. *Neurocomputing*, 52, 1–6. [https://doi.org/10.1016/s0925-2312\(02\)00738-5](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. (2001a). High Capacity Neural Networks for Familiarity Discrimination. *Journal of Computational Neuroscience*, 10(1), 5–23.
<https://doi.org/10.1023/A:1008925909305>
- Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001b). 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., Crupi, C., Pigott, S., Parrent, A., Wiebe, S., Janzen, L., & Köhler, S. (2010). Double dissociation of selective recollection and familiarity impairments following two different surgical treatments for temporal-lobe epilepsy. *Neuropsychologia*, 48(9), 2640–2647.
<https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2010.05.010>
- Boyle, L., Posani, L., Irfan, S., Siegelbaum, S. A., & Fusi, S. (2022). The geometry of hippocampal CA2 representations enables abstract coding of social familiarity and identity. *BioRxiv*, 2022.01.24.477361. <https://doi.org/10.1101/2022.01.24.477361>
- Brandt, K. R., Eysenck, M. W., Nielsen, M. K., & von Oertzen, T. J. (2016). Selective lesion to the entorhinal cortex leads to an impairment in familiarity but not recollection. *Brain and Cognition*, 104, 82–92. <https://doi.org/10.1016/J.BANDC.2016.02.005>
- Brown, A. S. (2003). A Review of the Déjà Vu Experience. *Psychological Bulletin*, 129(3), 394–413.
<https://doi.org/10.1037/0033-2909.129.3.394>

- 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](https://doi.org/10.1016/S0301-0082(98)00002-1)
- Cho, K., Kemp, N., Noel, J., Aggleton, J. P., Brown, M. W., & Bashir, Z. I. (2000). A new form of long-term depression in the perirhinal cortex. *Nature Neuroscience*, 3(2), 150–156.
<https://doi.org/10.1038/72093>
- 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>
- Cleary, A. M. (2008). Recognition Memory, Familiarity, and Déjà vu Experiences. *Current Directions in Psychological Science*, 17(5), 353–357. <https://doi.org/10.1111/j.1467-8721.2008.00605.x>
- Cleary, A. M., Ryals, A. J., & Nomi, J. S. (2009). Can déjà vu result from similarity to a prior experience? support for the similarity hypothesis of déjà vu. *Psychonomic Bulletin and Review*, 16(6), 1082–1088. <https://doi.org/10.3758/PBR.16.6.1082>
- Cowell, R. A. (2012). Computational models of perirhinal cortex function. *Hippocampus*, 22(10), 1952–1964. <https://doi.org/10.1002/hipo.22064>
- Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, 28(6), 923–938. <https://doi.org/10.3758/BF03209340>
- Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54(3), 180–187.
<https://doi.org/10.1037/h0040536>
- Defays, D., French, R. M., & Sougné, J. (1997). Apports de l'Intelligence Artificielle à la Psychologie. In J.-A. Rondal (Ed.), *Introduction à la psychologie scientifique* (Vol. 10, pp. 379–415). Labor.
- Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, 13(1), 1–21.
<https://doi.org/10.3758/BF03193807>
- Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2008). The Effects of Unitization on Familiarity-Based Source Memory: Testing a Behavioral Prediction Derived From Neuroimaging Data.

- Journal of Experimental Psychology: Learning Memory and Cognition*, 34(4), 730–740.
<https://doi.org/10.1037/0278-7393.34.4.730>
- Duke, D., Fiacconi, C. M., Köhler, S., & Clark, K. B. (2014). *Parallel effects of processing fluency and positive affect on familiarity-based recognition decisions for faces*.
<https://doi.org/10.3389/fpsyg.2014.00328>
- Düzel, E., Yonelinas, A., Mangun, G., Heinze, H.-J., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences*, 94(11), 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, 32, ii, 32–ii.
- Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, 30(1), 123–152.
<https://doi.org/10.1146/annurev.neuro.30.051606.094328>
- Elfman, K. W., Parks, C. M., & Yonelinas, A. P. (2008). Testing a Neurocomputational Model of Recollection, Familiarity, and Source Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, 34(4), 752–768. <https://doi.org/10.1037/0278-7393.34.4.752>
- Fiacconi, C. M., Owais, S., Peter, E. L., & Köhler, S. (2016). Knowing by heart: Visceral feedback shapes recognition memory judgments. *Journal of Experimental Psychology: General*, 145(5), 1–14. <https://doi.org/10.1037/xge0000164>
- French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 128–135. [https://doi.org/10.1016/S1364-6613\(99\)01294-2](https://doi.org/10.1016/S1364-6613(99)01294-2)
- French, R. M., Quinn, P. C., & Mareschal, D. (2001). Reversing Category Exclusivities in Infant Perceptual Categorization: Simulations and Data. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 23, 23.
- Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16(4), 309–313. <https://doi.org/10.3758/BF03197041>
- Geurten, M., Willems, S., Salmon, E., & Bastin, C. (2020). Fluency-based memory decisions in Alzheimer’s disease: A matter of source detection? *Neuropsychology*, 34(2), 176–185.
<https://doi.org/10.1037/neu0000605>
- Griffin, G., Holub, A. D., & Perona, P. (2007). *Caltech 256. Object category dataset: Caltech Technical Report*. (ImageNet). <https://doi.org/www.kaggle.com/datasets/jessicali9530/caltech256>

- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, 10(1), 14–23.
<https://doi.org/10.1016/J.TICS.2005.11.006>
- Grober, E., & Buschke, H. (1987). Genuine memory deficits in dementia. *Developmental Neuropsychology*, 3(1), 13–36. <https://doi.org/10.1080/87565648709540361>
- Grossberg, S. (1976). Biological Cybernetics Adaptive Pattern Classification and Universal Recoding: I. Parallel Development and Coding of Neural Feature Detectors. *Biol. Cybernetics*, 23, 121–134.
<https://doi.org/10.1007/BF00344744>
- Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *European Conference on Computer Vision*, 87–102.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.
- Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation* (1st ed.). CRC Press.
- Hintzman, D. L., Caulton, D. A., & Levitin, D. J. (1998). Retrieval dynamics in recognition and list discrimination: Further evidence of separate processes of familiarity and recall. *Memory & Cognition*, 26(3), 449–462. <https://doi.org/10.3758/BF03201155>
- Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? [Article]. *Journal of Experimental Psychology.*, 18(4), 667–680.
<https://doi.org/10.1037/0278-7393.18.4.667>
- Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus*, 12(3), 341–351.
<https://doi.org/10.1002/hipo.10011>
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558.
<https://doi.org/10.1073/pnas.79.8.2554>

- Humphreys, G. W., & Riddoch, M. J. (2006). Features, objects, action: The cognitive neuropsychology of visual object processing, 1984–2004. *Cognitive Neuropsychology*, 23(1), 156–183. <https://doi.org/10.1080/02643290542000030>
- Ito, M. (1989). Long-Term Depression. *Annual Review of Neuroscience*, 12(1), 85–102. <https://doi.org/10.1146/annurev.ne.12.030189.000505>
- 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](https://doi.org/10.1016/0749-596X(91)90025-F)
- Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. [Article]. *Journal of Experimental Psychology.*, 110(3), 306–340. <https://doi.org/10.1037/0096-3445.110.3.306>
- Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. [Article]. *Journal of Experimental Psychology.*, 122(2), 139–154. <https://doi.org/10.1037/0096-3445.122.2.139>
- Jacoby, L. L., & Whitehouse, K. (1989). An Illusion of Memory: False Recognition Influenced by Unconscious Perception. *Journal of Experimental Psychology: General*, 118(2), 126–135. <https://doi.org/10.1037/0096-3445.118.2.126>
- Ji-An, L., Stefanini, F., Benna, M. K., & Fusi, S. (2022). Face familiarity detection with complex synapses. *BioRxiv*, 854059. <https://doi.org/10.1101/854059>
- 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>
- 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>
- Klimesch, W., Doppelmayr, M., Yonelinas, A., Kroll, N. E. A., Lazzara, M., Röhms, D., & Gruber, W. (2001). Theta synchronization during episodic retrieval: neural correlates of conscious awareness. *Cognitive Brain Research*, 12(1), 33–38. [https://doi.org/10.1016/S0926-6410\(01\)00024-6](https://doi.org/10.1016/S0926-6410(01)00024-6)
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1097–1105). Curran

Associates, Inc.

<https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>

- le Cun, Y. (2019). *Quand la machine apprend: la révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob.
- Logie, R. H. (1996). The seven ages of working memory. *Working Memory and Human Cognition*, 31–65.
- 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>
- Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, 36(5), 635–645. <https://doi.org/10.1037/0012-1649.36.5.635>
- Martin, C. B., McLean, D. A., O'Neil, E. B., & Köhler, S. (2013). Distinct familiarity-based response patterns for faces and buildings in perirhinal and parahippocampal cortex. *Journal of Neuroscience*, 33(26), 10915–10923. <https://doi.org/10.1523/JNEUROSCI.0126-13.2013>
- Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the Hippocampus. *Hippocampus*, 12(3), 325–340. <https://doi.org/10.1002/hipo.1111>
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, 102(3), 419–457. <https://doi.org/10.1037/0033-295X.102.3.419>
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C), 109–165. [https://doi.org/10.1016/S0079-7421\(08\)60536-8](https://doi.org/10.1016/S0079-7421(08)60536-8)
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>
- Merkow, M. N., Burke, J. F., & Kahana, M. J. (2015). The human hippocampus contributes to both the recollection and familiarity components of recognition memory. *Proceedings of the National Academy of Sciences*, 112(46), 14378–14383. <https://doi.org/10.1073/pnas.1513145112>
- Meyer, T., & Rust, N. C. (2018). *Single-exposure visual memory judgments are reflected in inferotemporal cortex*. <https://doi.org/10.7554/eLife.32259.001>

- Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, 62(6), 1198–1215.
<https://doi.org/10.1080/17470210802391599>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. [Article]. *The Psychological Review.*, 63(2), 81–97.
<https://doi.org/10.1037/h0043158>
- Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin UK.
- Montaldi, D., & Mayes, A. R. (2010). The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. *Hippocampus*, 20(11), 1291–1314.
<https://doi.org/https://doi.org/10.1002/hipo.20853>
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482. <https://doi.org/10.1037/h0045106>
- Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, 20(11), 1217–1227.
<https://doi.org/10.1002/hipo.20855>
- 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>
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). *Deep Face Recognition*.
- Quamme, J. R., Yonelinas, A. P., & Norman, K. A. (2007). Effect of unitization on associative recognition in amnesia. *Hippocampus*, 17(3), 192–200. <https://doi.org/10.1002/hipo.20257>
- Ranganath, C. (2010). A unified framework for the functional organization of the medial temporal lobes and the phenomenology of episodic memory. *Hippocampus*, 20(11), 1263–1290.
<https://doi.org/https://doi.org/10.1002/hipo.20852>
- Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. [Article]. *The Psychological Review.*, 97(2), 285–308.
<https://doi.org/10.1037/0033-295X.97.2.285>

- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386. <https://doi.org/10.1037/h0042519>
- Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, 28(6), 907–922. <https://doi.org/10.3758/BF03209339>
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing*. The MIT Press. <https://doi.org/10.7551/mitpress/5236.001.0001>
- Scalici, F., Caltagirone, C., & Carlesimo, G. A. (2017). The contribution of different prefrontal cortex regions to recollection and familiarity: a review of fMRI data. *Neuroscience & Biobehavioral Reviews*, 83, 240–251. <https://doi.org/10.1016/J.NEUBIOREV.2017.10.017>
- Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annu. Rev. Psychol*, 49(1), 289–318. <https://doi.org/10.1146/annurev.psych.49.1.289>
- Schulz, A., Miehl, C., Berry, M. J., & Gjorgjieva, J. (2021). The generation of cortical novelty responses through inhibitory plasticity. *ELife*, 10. <https://doi.org/10.7554/eLife.65309>
- Shapiro, S. C. (1992). Encyclopedia of artificial intelligence second edition. *New Jersey: A Wiley Interscience Publication*.
- She, L., Benna, M. K., Shi, Y., Fusi, S., & Tsao, D. Y. (2021). The neural code for face memory. *BioRxiv*, 2021.03.12.435023. <https://doi.org/10.1101/2021.03.12.435023>
- Sougné, J. (2002). Short Term Memory in a Network of Spiking Neurons. In *Connectionist Models of Cognition and Perception* (pp. 131–142). https://doi.org/10.1142/9789812777256_0011
- Sougné, J., & French, R. (2001, June 23). Synfire Chains and Catastrophic Interference. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- 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>
- Squire, L. R., & Wixted, J. T. (2011). The Cognitive Neuroscience of Human Memory Since H.M. *Annual Review of Neuroscience*, 34(1), 259–288. <https://doi.org/10.1146/annurev-neuro-061010-113720>
- Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, 8(11), 872–883. <https://doi.org/10.1038/nrn2154>

- Squire, L. R., & Zola, S. M. (1998). Episodic Memory, Semantic Memory, and Amnesia. *Hippocampus*, 8(3), 205–211. [https://doi.org/10.1002/\(SICI\)1098-1063\(1998\)8:3<205::AID-HIPO3>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1098-1063(1998)8:3<205::AID-HIPO3>3.0.CO;2-I)
- Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25, 207–222. <https://doi.org/10.1080/14640747308400340>
- Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A Cortical Region Consisting Entirely of Face-Selective Cells. *Science*, 311(5761), 670–674. <https://doi.org/10.1126/science.11119983>
- Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, 26(1), 1. <https://doi.org/10.1037/h0080017>
- Tulving, E., & Markowitsch, H. J. (1998). Episodic and Declarative Memory: Role of the Hippocampus. *Hippocampus*, 8, 198–204. [https://doi.org/10.1002/\(SICI\)1098-1063\(1998\)8:3<198::AID-HIPO2>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1098-1063(1998)8:3<198::AID-HIPO2>3.0.CO;2-G)
- 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>
- Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The Hippocampus Supports both the Recollection and the Familiarity Components of Recognition Memory. *Neuron*, 49(3), 459–466. <https://doi.org/10.1016/J.NEURON.2005.12.020>
- Weschler, D. (2001). Echelle clinique de memoire de Weschler MEM III (WMS-III). *Les Editions Du Centre de Psychologie Appliquee, Paris*.
- Westerberg, C. E., Paller, K. A., Weintraub, S., Mesulam, M. M., Mayes, A. R., Holdstock, J. S., & Reber, P. J. (2006). When memory does not fail: Familiarity-based recognition in mild cognitive impairment and Alzheimer's disease. *Neuropsychology*, 20(2), 193–205. <https://doi.org/10.1037/0894-4105.20.2.193>
- Whittlesea, B. W. A. (1993). Illusions of Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(6), 1235–1253. <https://doi.org/10.1037/0278-7393.19.6.1235>
- Wolk, D. A., Dunfee, K. L., Dickerson, B. C., Aizenstein, H. J., & Dekosky, S. T. (2011). A Medial Temporal Lobe Division of Labor: Insights From Memory in Aging and Early Alzheimer Disease. *Hippocampus*, 21(5), 461–466. <https://doi.org/10.1002/hipo.20779>

- Wolk, D. A., Gold, C. A., Signoff, E. D., & Budson, A. E. (2009). Discrimination and reliance on conceptual fluency cues are inversely related in patients with mild Alzheimer's disease. *Neuropsychologia*, 47(8–9), 1865–1872. <https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2009.02.029>
- Wolk, D. A., Schacter, D. L., Berman, A. R., Holcomb, P. J., Daffner, K. R., & Budson, A. E. (2005). Patients with mild Alzheimer's disease attribute conceptual fluency to prior experience. *Neuropsychologia*, 43(11), 1662–1672. <https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2005.01.007>
- Wolk, D. A., Signoff, E. D., & DeKosky, S. T. (2008). Recollection and familiarity in amnesic 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](https://doi.org/10.1016/S0028-3908(98)00030-6)
- Yonelinas, A., Kroll, N. E., Dobbins, G., & Soltani, M. (1999). Recognition memory for faces: When familiarity supports associative recognition judgments. *Psychonomic Bulletin & Review*, 6(4), 654–661. <https://doi.org/10.3758/BF03212975>
- 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., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-Detection, Threshold, and Dual-Process Models of Recognition Memory: ROCs and Conscious Recollection. *Consciousness and Cognition*, 5(4), 418–441. <https://doi.org/10.1006/CCOG.1996.0026>
- Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of Processes in Recognition Memory: Effects of Interference and of Response Speed. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 48(4), 516–535. <https://doi.org/10.1037/1196-1961.48.4.516>

- Yonelinas, A. P., & Parks, C. M. (2007). Receiver Operating Characteristics (ROCs) in Recognition Memory: A Review. *Psychological Bulletin*, 133(5), 800–832. <https://doi.org/10.1037/0033-2909.133.5.800>
- Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., & Poo, M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395(6697), 37–44. <https://doi.org/10.1038/25665>
- Zhang, W., Sun, J., & Tang, X. (2008). Cat head detection-how to effectively exploit shape and texture features. *European Conference on Computer Vision*, 802–816.
- Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., & Gentner, T. Q. (2020). Five Ways in Which Computational Modeling Can Help Advance Cognitive Science: Lessons From Artificial Grammar Learning. *Topics in Cognitive Science*, 12(3), 925–941. <https://doi.org/10.1111/tops.12474>

ANNEXES

Annexe 1. Extraction des caractéristiques d'une image avec AlexNet

Dans le modèle FaRe, le module d'extraction de caractéristiques est un RPC du nom d'AlexNet (Krizhevsky et al., 2012) préalablement entraîné sur 1.3 millions de photographies haute résolution d'images naturelles. Cette architecture permet la classification d'images dans 1000 catégories différentes. Elle est constituée de 25 couches, dont les 19 premières sont une combinaison de couches de convolution, de *pooling* et des modules de normalisation permettant d'analyser une image et d'en définir ses caractéristiques selon différents degrés de complexité. La 20ème couche, appelée *fc7*, est *fully-connected* et représente l'intégration des caractéristiques extraites par les couches précédentes. Les couches suivantes constituent le classifieur et reçoivent les *features* obtenus par la couche *fc7* pour classifier l'image grâce à un algorithme de rétropropagation.

Le vecteur obtenu à la couche *fc7* pour une image donnée est utilisé en entrées dans le module de mémoire. Après être passé dans AlexNet, le vecteur de taille 4096 correspondant aux caractéristiques de l'image est récolté, avant d'être normalisé avec une moyenne de 0 et un écart-type de 1. Le vecteur normalisé constitue les entrées pour le module de mémoire. La **Figure A1** montre les histogrammes des valeurs de sortie obtenues après cette étape.

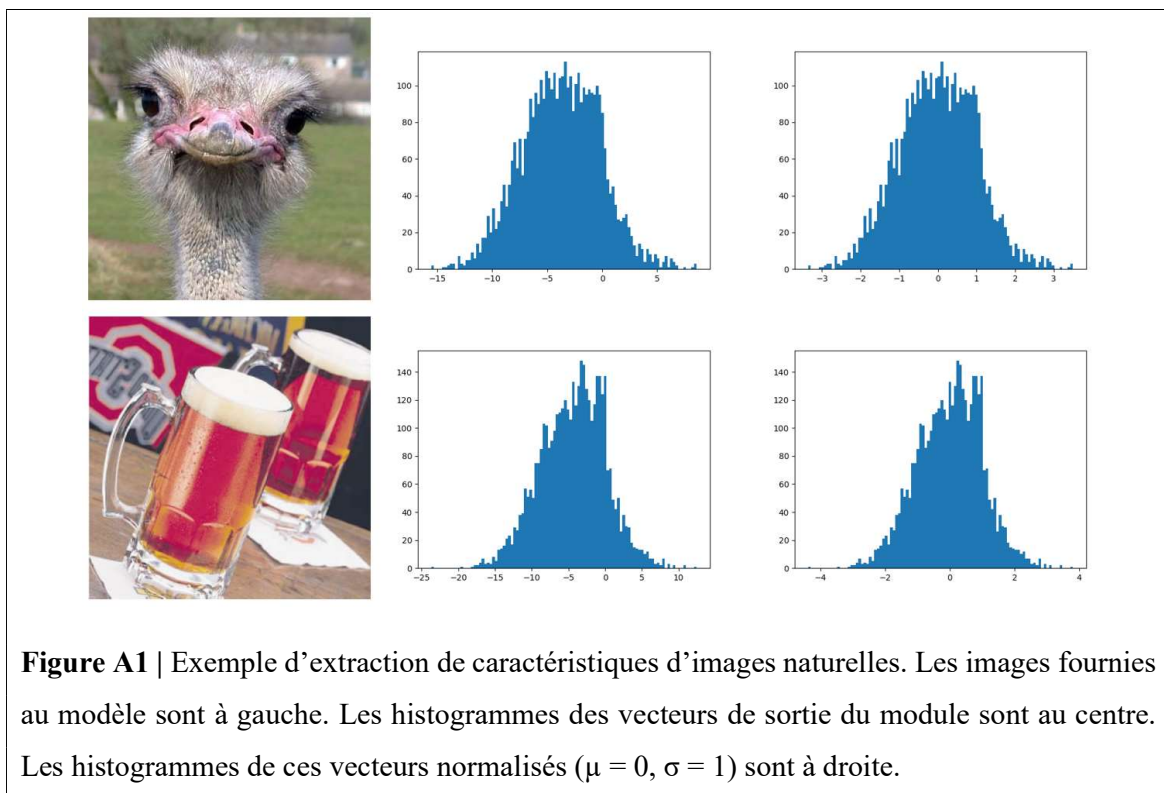


Figure A1 | Exemple d'extraction de caractéristiques d'images naturelles. Les images fournies au modèle sont à gauche. Les histogrammes des vecteurs de sortie du module sont au centre. Les histogrammes de ces vecteurs normalisés ($\mu = 0$, $\sigma = 1$) sont à droite.

Annexe 2. Module de mémoire Anti-Hebbien

Le module de mémoire anti-Hebbien est un réseau de neurones *feedforward* composé de deux couches *fully-connected*. Ce réseau apprend selon une règle d'apprentissage anti-Hebbienne, adaptée à des vecteurs d'entrées constitués de nombres réels plutôt que de valeurs binaires telles que dans les modèles précédents (Androulidakis et al., 2008 ; Bogacz & Brown, 2003a). Par ailleurs, et comme déjà mentionné précédemment, ce type d'apprentissage est biologiquement plausible (Bogacz & Brown, 2003a) et permet la modélisation d'un grand nombre de données expérimentales (Androulidakis et al., 2008 ; Kazanovich & Borisyuk, 2021).

La couche d'entrée contient n neurones à l'instar de la couche de sortie, également composée de m neurones de nouveauté. Les deux couches sont entièrement connectées l'une à l'autre par des poids, qui sont distribués aléatoirement entre -1 et 1 lors de l'initialisation du modèle. Le potentiel membranaire (h_j) des neurones de nouveauté est déterminé par la formule suivante :

$$h_j = \sum_{i=1}^n w_{ij} x_i, \quad j = 1, \dots, m, \quad i = 1, \dots, n$$

où x_i correspond au vecteur de caractéristiques de l'image X, produites à la couche fc7 d'AlexNet et après normalisation, et w_{ij} correspond aux poids entre les neurones des couches d'entrée et de sortie.

La couche de sortie fonctionne grâce à un mécanisme de compétition et d'inhibition *m/2-winners*. C'est-à-dire que seule la moitié des neurones avec les plus hauts potentiels membranaires sont considérés comme étant actifs. Les autres neurones de nouveauté sont considérés comme étant au repos et ne participent pas à la modification des poids durant l'apprentissage. Ainsi, les neurones de nouveauté dont le potentiel membranaire est $>$ à la médiane de l'activité des sorties prennent la valeur $y_j = 1$; ceux dont le potentiel est $\leq$ à la médiane prennent la valeur $y_j = 0$. Les poids des neurones de sortie sont ensuite modifiés lors de l'entraînement grâce à la règle d'apprentissage anti-Hebbien suivante :

$$w_{ij} = w_{ij} - \eta x_i y_j$$

où $\eta > 0$ et correspond à la *learning rate*, w_{ij} aux poids, x_i au vecteur de caractéristiques de l'image X et y_j au vecteur de neurones de nouveauté.

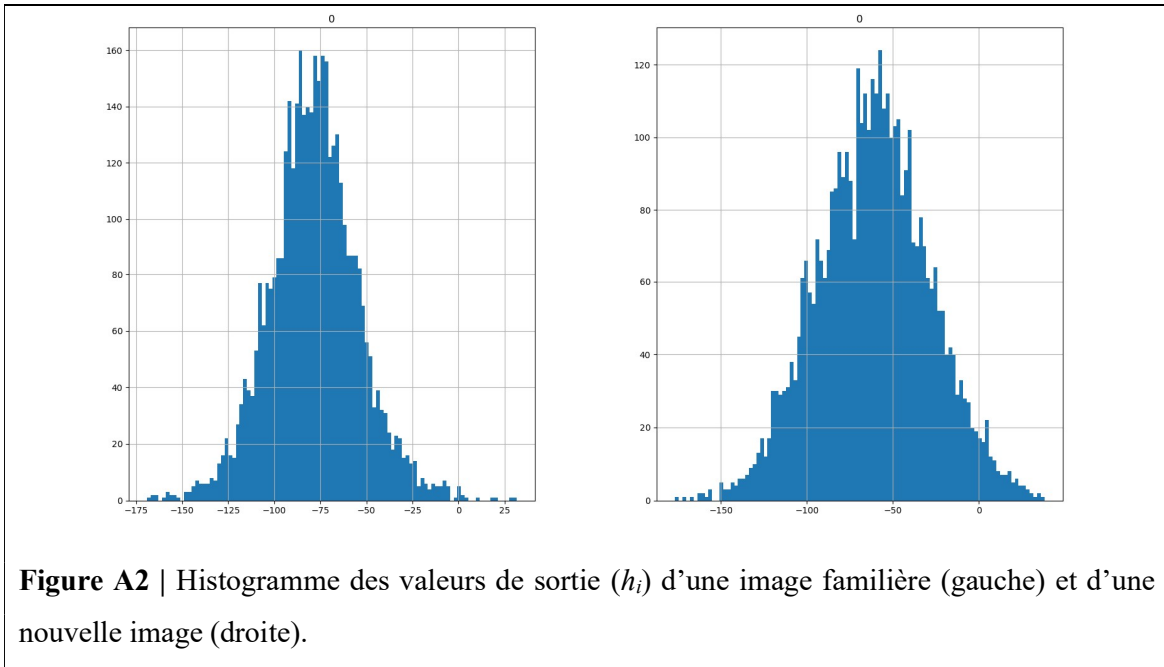
Cet apprentissage anti-Hebbien permet, durant la phase d'entraînement, de modifier les poids entre les neurones de sorte que l'activité moyenne de la couche de sortie diminue. Lors de la phase d'entraînement, cette modification de poids est implémentée sur 1 epoch pour chaque image X du *training set*. Lors de la phase de test, l'activité moyenne d'une paire d'images (X, Z) , où X est une image étudiée lors de l'entraînement et Z est une nouvelle image, est calculée via les formules suivantes :

$$d(X) = \frac{1}{m} \left(\sum_{j \in M_1} \sum_{i=1}^n w_{ij} x_i - \sum_{j \in M_2} \sum_{i=1}^n w_{ij} x_i \right)$$

$$d(Z) = \frac{1}{m} \left(\sum_{j \in M_1} \sum_{i=1}^n w_{ij} z_i - \sum_{j \in M_2} \sum_{i=1}^n w_{ij} z_i \right)$$

où M_1 et M_2 sont respectivement les $m/2$ -winners et losers dans la couche de sortie du réseau.

Etant donné qu'une image familière est supposée produire moins d'activité qu'une nouvelle image, une décision correcte de familiarité aura lieu si $d(X) < d(Z)$. Si ce n'est pas le cas, une fausse reconnaissance est encodée. La **Figure A2** représente les histogrammes des valeurs de l'activité d'une image familière par rapport à une nouvelle image.



Les propriétés initiales du module de mémoire pour les simulations avec AlexNet sont les suivantes : $n = 4096$ neurones, $m = 4096$ neurones et $\eta = 0,01$.

Annexe 3. Implémentation d'un test de reconnaissance Oui/Non

Les détails méthodologiques de la modélisation d'une tâche de reconnaissance Oui/Non sont décrits ci-dessous.

Dataset

Le *dataset* utilisé pour la simulation avec les images de catégories différentes provient de la base de données « Caltech 256 » (Griffin et al., 2007). Celui pour les images d'une même catégorie est le « Cat Dataset » utilisé dans la Simulation 3 (W. Zhang et al., 2008).

Méthodologie

Avant de simuler le test de reconnaissance Oui/Non, nous avons d'abord dû *hardcoder* un seuil au-delà duquel le modèle considère une image comme familière ou non. Pour ce faire, nous avons entraîné le modèle sur 400 images. Après l'entraînement, nous avons calculé le score de familiarité pour ces images ainsi que pour 400 nouvelles images avec la formule suivante :

$$d(X) = \frac{1}{m} \left(\sum_{j \in M_1} \sum_{i=1}^n w_{ij} x_i - \sum_{j \in M_2} \sum_{i=1}^n w_{ij} x_i \right)$$

Où M_1 et M_2 sont respectivement les $m/2$ -winners et losers dans la couche de sortie du réseau.

Nous avons ensuite calculé la distribution des scores de familiarité pour l'ensemble de ces images. Le seuil de familiarité a été placé approximativement à l'intersection entre les deux courbes (**Figure A3**). Pour cette simulation, le seuil a été positionné à 56000.

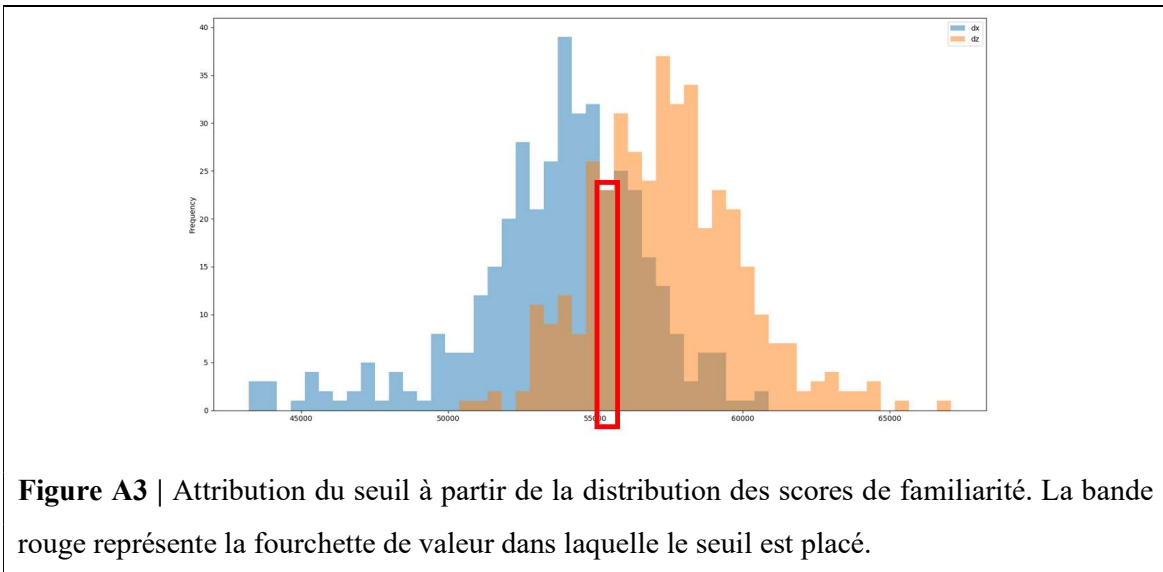


Figure A3 | Attribution du seuil à partir de la distribution des scores de familiarité. La bande rouge représente la fourchette de valeur dans laquelle le seuil est placé.

Nous avons ensuite récupéré aléatoirement 20 images sur lesquels le modèle a été entraîné au préalable et 20 nouvelles images jamais vues auparavant. La formule précédente a été appliquée sur les 40 images et la valeur de $d(X)$ obtenue pour chaque image a été comparée à la valeur seuil. Si $d(X)$ est inférieur au seuil, alors l'image est considérée comme familière. Dans le cas contraire, l'image est considérée comme nouvelle (**Figure A4**).

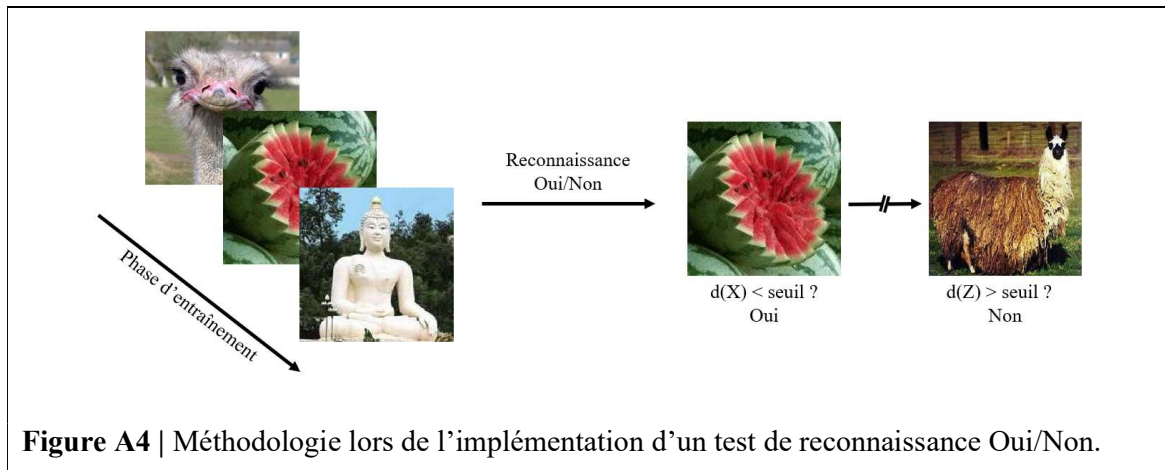


Figure A4 | Méthodologie lors de l'implémentation d'un test de reconnaissance Oui/Non.

Résultats

Les résultats sont présentés sur la **Figure A5**, à côté de ceux obtenus à la Simulation 3.

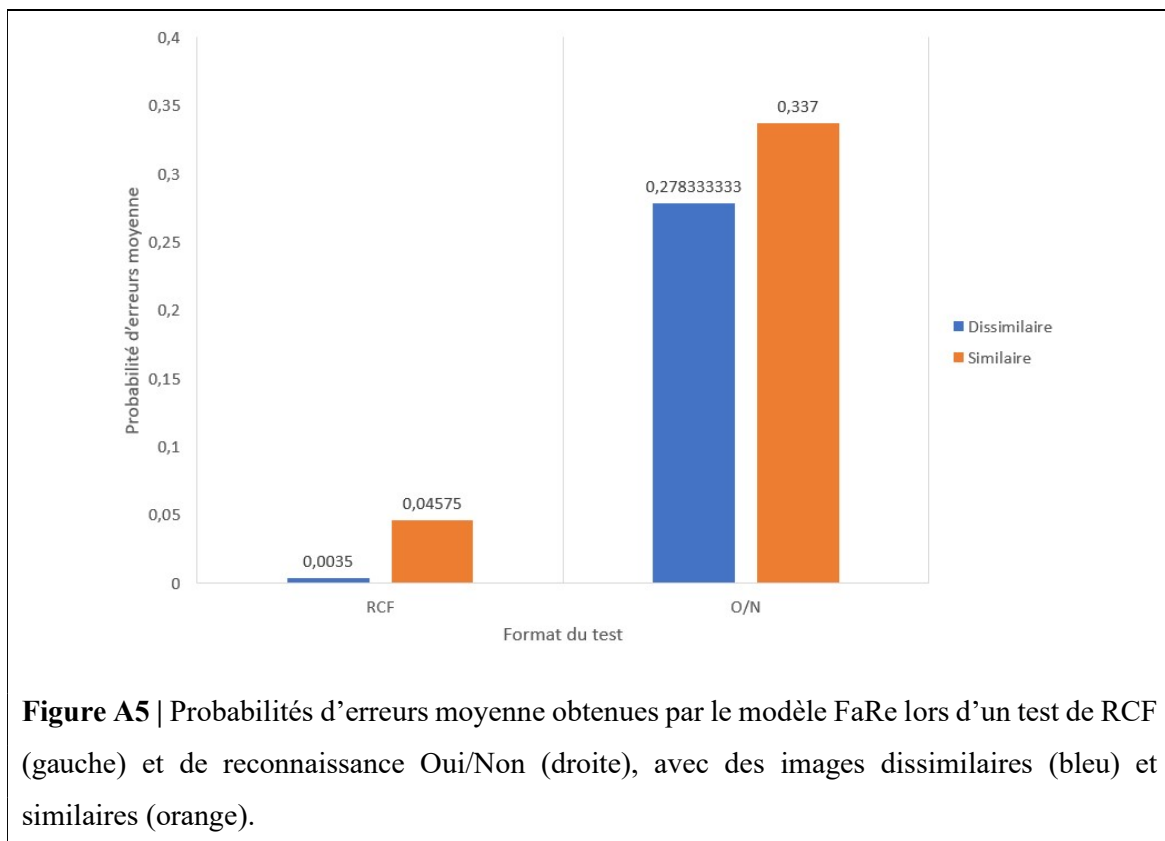


Figure A5 | Probabilités d'erreurs moyenne obtenues par le modèle FaRe lors d'un test de RCF (gauche) et de reconnaissance Oui/Non (droite), avec des images dissimilaires (bleu) et similaires (orange).

ABSTRACT

Les théories actuelles postulent que la reconnaissance peut s'effectuer selon deux processus indépendants mais qui agissent en parallèle : la recollection et la familiarité (Yonelinas, 2002). Par le passé, plusieurs modèles computationnels ont tenté de reproduire artificiellement le sentiment de familiarité, afin d'apporter du soutien à ces théories des deux processus et ainsi comprendre les mécanismes qui sous-tendent la familiarité (Bogacz & Brown, 2003b ; Kazanovich & Borisyuk, 2021).

Dans ce mémoire, nous avons programmé un modèle de neurones artificiels pour la reconnaissance par familiarité sur des images naturelles dans le cortex périrhinal. Ce modèle a été conçu comme un système en deux étapes. Dans la première étape, nous utilisons ResNet50, un réseau profond convolutif (RPC) pré-entraîné pour extraire les caractéristiques d'une image (He et al., 2016). Dans la seconde étape, un réseau de neurones deux-couches à propagation avant, utilisant une règle d'apprentissage Hebbienne (Hebb, 1949), est utilisé pour la décision de familiarité à propos d'une image.

Nous avons implémenté une tâche de reconnaissance à choix forcés (RCF) et réalisé quatre simulations afin de tester les capacités de notre modèle. Premièrement, le modèle montre une capacité de mémoire allant jusqu'à 40 images. Deuxièmement, il présente un effet de récence lorsque le nombre d'images qui lui est présentée à l'entraînement ne dépasse pas sa capacité mnésique. Troisièmement, le modèle performe significativement moins bien lorsque les images présentées lors de la tâche de reconnaissance appartiennent toutes à la même catégorie sémantique, montrant ainsi un effet de la similarité sur ses performances. Finalement, le modèle parvient à reconnaître des visages lorsque ceux-ci sont présentés dans différentes positions avec une précision supérieure à 75%.