# Modèles connexionnistes de la mémoire

Bernard Victorri

▶ **To cite this version:**

Bernard Victorri. Modèles connexionnistes de la mémoire. F. Eustache, B. Lechevalier, F. Viader. La Mémoire – Neuropsychologie clinique et modèles cognitifs, De Boeck, pp.371-387, 1995. halshs-00138763

**HAL Id: halshs-00138763**
**https://shs.hal.science/halshs-00138763**

Submitted on 27 Mar 2007

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Modèles connexionnistes de la mémoire

## Introduction[1]

Depuis une dizaine d'années, avec l'apparition du connexionnisme, on assiste au développement d'une voie originale de recherche dans l'étude des phénomènes cognitifs. En effet, les modèles connexionnistes, basés sur l'utilisation de ce que l'on appelle les réseaux de « neurones formels », permettent de simuler, de manière certes très simpliste, des comportements du type de ceux que l'on observe en psychologie expérimentale, à l'aide de systèmes dont l'organisation paraît proche, à un certain niveau d'abstraction en tout cas, des systèmes physico-chimiques que nous révèle la neurophysiologie. Ainsi ces modèles semblent à terme être en mesure de combler le fossé entre fonctionnement du cerveau et fonctionnement de l'esprit (*brain and mind* dans la terminologie anglo-saxone), en proposant des mécanismes « réalistes » du point de vue de la physiologie capables d'expliquer certaines caractéristiques de la cognition humaine. D'où l'intérêt croissant pour cette activité de modélisation, autant de la part des spécialistes de l'intelligence artificielle, qui espèrent trouver là un moyen de simuler des aspects de l'intelligence humaine particulièrement réfractaires aux méthodes classiques fondées sur le calcul symbolique, que de la part des spécialistes des sciences cognitives, qui se heurtent de front au problème des relations entre les processus psychiques et les processus physiologiques qui leur servent de substrat. C'est ainsi que le connexionnisme peut contribuer à renouveler certaines problématiques en neuropsychologie : par exemple, Shallice (1991) montre que les résultats obtenus avec certains modèles connexionnistes pourraient remettre en cause l'interprétation d'un grand nombre de données neuropsychologiques, comme les « doubles dissociations », qui ont été jusqu'à présent considérées comme des preuves indiscutables du modularisme.

La mémoire est l'un des domaines pour lesquels la modélisation connexionniste semble des plus prometteuses : d'une part, les caractéristiques de la mémoire humaine sont très éloignées des représentations classiques qu'en donne l'ordinateur, et d'autre part, les connaissances actuelles sur le cerveau sont encore loin de donner une vision claire et cohérente des mécanismes qui sous-tendent ce phénomène. Le connexionnisme paraît donc tout indiqué pour tenter de mieux comprendre comment des processus neuronaux peuvent permettre les activités de mémorisation et de rappel qui constituent une composante essentielle de la plupart des activités cognitives humaines. Le but de cet article est d'essayer d'évaluer dans quelle mesure cette voie de recherche est susceptible de répondre effectivement à cette attente, et de délimiter de manière plus précise la place qu'elle peut tenir dans l'étude de la mémoire humaine.

## 1. Modélisation

### 1.1. Principe

Nous nous appuierons essentiellement sur les travaux fondateurs de McClelland et Rumelhart (1986a) qui ont joué un rôle important en popularisant largement aussi bien les techniques utilisées que les enjeux de ces recherches. Depuis, d'autres modèles ont bien sûr été proposés, mais les principes fondamentaux, qui seuls nous importeront ici, sont restés inchangés.

Le point de départ de McClelland et Rumelhart consiste à considérer que le support matériel de la mémoire est constitué d'une série de modules, construits sur un même modèle, formant un vaste réseau : chaque module reçoit en entrée l'information d'un certain nombre d'entre eux, et envoie en sortie l'information traitée à d'autres, les boucles n'étant pas exclues. Quelques uns de ces modules peuvent être directement reliés au monde extérieur, soit par leurs entrées (modules « perceptifs »), soit par leurs sorties (modules « moteurs »). Un module est donc conçu comme une unité de traitement de

[1] Une version préliminaire abrégée de cet article est parue sous le titre "Mémoire et connexionnisme», dans *Les Cahiers de la Maison de la Recherche en Sciences Humaines de Caen,* n° 2, Université de Caen, 1994

l'information, capable de combiner des informations en provenance de plusieurs sources, et servant à son tour de source pour des traitements ultérieurs (cf. fig. 1). Le but de McClelland et Rumelhart est de montrer qu'en supposant une organisation interne somme toute assez simple de chacun de ces modules, ils peuvent être le siège de phénomènes de mémorisation et de rappel, en un sens qu'il faut bien sûr soigneusement préciser.

Chaque module est lui-même constitué d'unités, ou neurones formels, entièrement interconnectées, qui sont caractérisées chacune par une variable d'état, comprise entre –1 et +1, que l'on appelle état d'activité de l'unité. Cet état évolue dans le temps, sous l'influence de l'état des autre unités et des entrées du module. Les connexions entre unités d'un même module sont caractérisées chacune par un nombre réel (compris entre $-\infty$ et $+\infty$) que l'on appelle le poids de la connexion. La modélisation repose sur les trois principes de base suivants :

- Un état « mental » est représenté dans le modèle par la donnée d'un pattern d'activité de toutes les unités des différents modules du système.
- La mémorisation d'un état mental correspond à des modifications des valeurs des poids des liaisons entre unités.
- Le rappel d'un état mémorisé correspond au rétablissement d'un pattern d'activité antérieur.

L'image que donne le modèle peut donc se résumer de la façon suivante. Supposons que le système soit soumis à un environnement donné. Le pattern d'activité des unités va évoluer dans le temps, et, éventuellement, si l'environnement ne change pas pendant un temps suffisant, il va se stabiliser dans un état qui correspond à la perception des différents éléments de cet environnement. Des modifications de poids interviennent alors : ces modifications ne changent pas l'état d'activité présent (la perception reste la même), mais elles transforment le système, puisque celui-ci réagira dès lors de manière différente à de nouvelles situations. En particulier, si ces modifications conduisent certains modules à se retrouver plus tard dans ce même état, alors que l'environnement n'est plus le même, on interprétera cela comme un rappel de la situation antérieure, et l'on considérera que les modifications de poids correspondent à une trace mnésique de cette situation.

Il faut imaginer un système constitué de centaines, voire de milliers de modules en interaction, chaque module comportant lui-même un très grand nombre d'unités, par exemple de l'ordre de plusieurs millions. On conçoit alors la complexité des opérations qui pourraient être modélisées : certains modules pourraient traiter l'information entrante, sous ces différents aspects, tandis qu'en parallèle, d'autres modules pourraient être plus particulièrement dévolus à l'activité de mémorisation et de rappel, soit directement à partir des données « sensorielles » entrantes, soit à partir d'une information déjà traitée, permettant ainsi la coexistence de différents types de phénomènes mnésiques, y compris des mémoires « en cascade », l'information issue de ces modules de mémoire pouvant être à son tour combinée avec l'information entrante, etc. Bien entendu, il est hors de question pour le moment de construire des architectures aussi complexes : aussi McClelland et Rumelhart se limitent-ils à l'étude d'un seul module, et pour leurs simulations informatiques ce module est réduit à quelques dizaines d'unités, ce qui évidemment n'a plus rien à voir avec les conditions « réalistes » que nous venons de présenter (les questions d'ordre de grandeur ont une importance décisive en informatique : nous aurons l'occasion d'y revenir), mais qui suffit à mettre en évidence un certain nombre de propriétés intéressantes de ce type de système.

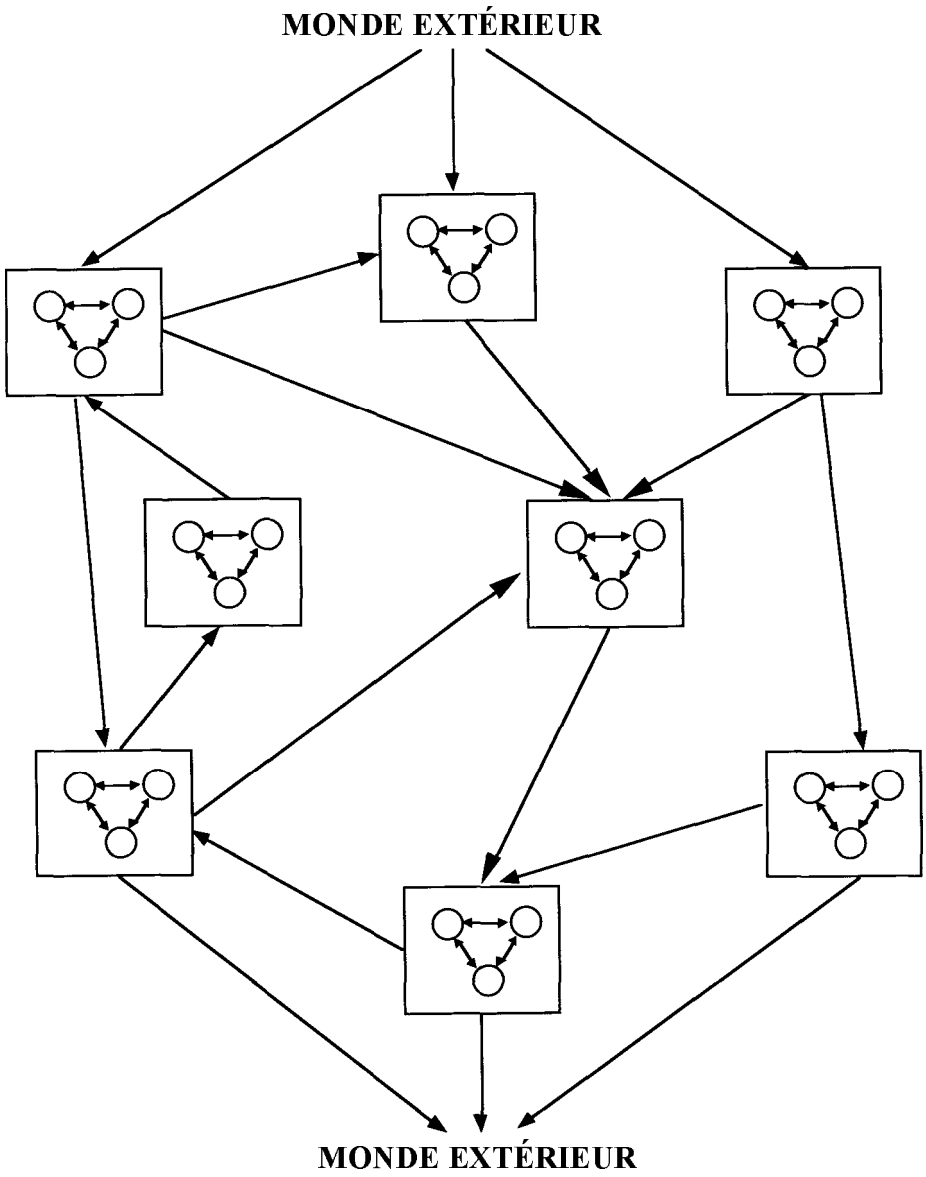


Figure 1 : Organisation générale du modèle

## 1.2. Fonctionnement

Si l'on se limite donc à un seul module, on a donc affaire à un réseau entièrement interconnecté, chaque unité étant reliée à toutes les autres (sauf elle-même), et recevant en plus une influence externe, qui correspond à une somme des entrées provenant des autres modules du système (et éventuellement du monde extérieur). A chaque instant, son activité est modifiée en fonction de la valeur de ces influences internes et externes. Si l'on appelle $a_i$ l'activité de la i-ème unité et $w_{ij}$ le poids de la liaison qui provient de la j-ème unité (cf. fig. 2), l'influence interne qu'elle reçoit, notée $i_i$ est donnée par une simple combinaison linéaire :

$$i_i = \sum_j w_{ij} a_j$$

Si l'on appelle $e_i$ la somme des entrées externes, la valeur de l'entrée totale, notée $A_i$, est donc simplement :

$$A_i - i_i + e_i$$

La loi qu'utilisent McClelland et Rumelhart pour calculer la modification de l'activité entre l'instant t et l'instant t+ 1, que nous noterons $\Delta a_i$, s'écrit alors :

si $A_i > 0$, alors $\Delta a_i = k A_i (1 - a_i) - h a_i$

si $A_i = 0$, alors $\Delta a_i = - h a_i$

si $A_i < 0$, alors $\Delta a_i = k A_i (1 + a_i) - h a_i$

Autrement dit, l'activité augmente et se rapproche de +1 si l'entrée totale est suffisamment positive (excitation), elle se rapproche de 0 si l'entrée totale est nulle ou très faible, et elle baisse et se rapproche de –1 si l'entrée totale est suffisamment négative (inhibition). Les constantes k et h sont identiques pour toutes les unités, et sont choisies une fois pour toutes : k exprime l'excitabilité plus ou moins grande des unités, et h le déclin plus ou moins rapide de l'activité vers la valeur de repos 0. D'autres lois ont été proposées, mais elles se conforment à cette même description qualitative.

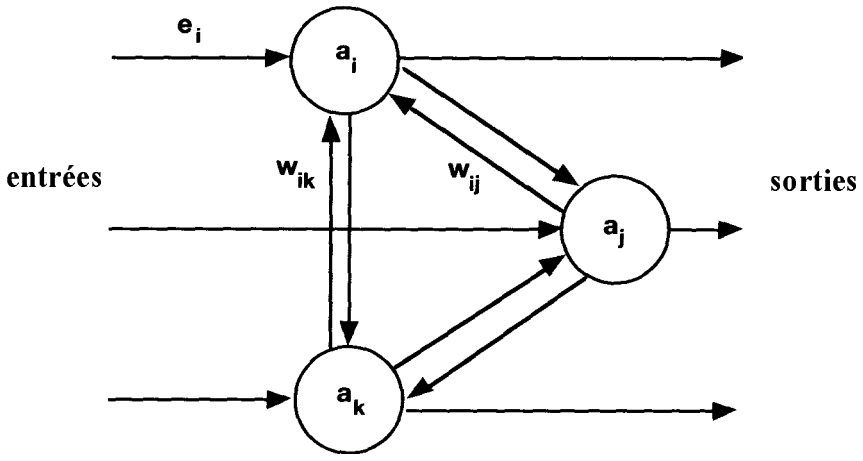


Figure 2 : Organisation interne d'un module

Supposons alors que les entrées externes restent constantes pendant un temps suffisamment long (en pratique, quelques dizaines d'unités de temps). Les activités des unités vont évoluer selon la loi précédente et vont finir par se stabiliser dans un état d'équilibre, que l'on appelle un attracteur de la dynamique induite par les entrées sur le système constitué par les unités du module. Ainsi à chaque pattern d'entrée $\{e_i\}$ correspond un pattern stable d'activité$\{a_i\}$ , qui représente en quelque sorte la réponse du module à ce pattern d'entrée. Le processus de mémorisation va consister alors à modifier les poids internes du réseau de façon à renforcer la probabilité d'apparition de cette réponse : plus précisément, de manière à ce que le système se restabilise dans l'état $\{a_i\}$ si l'on présente un pattern d'entrée qui a un certain nombre de points communs avec $\{e_i\}$. Mathématiquement parlant, le pattern $\{a_i\}$ sera mémorisé quand le bassin d'attracteur associé à $\{a_i\}$ (l'ensemble des états qui convergent vers cet attracteur) sera suffisamment large pour « attirer » tous les patterns d'entrée proches de $\{e_i\}$.

Pour obtenir ce résultat, la modification des poids obéit au principe qualitatif suivant : les poids de liens qui arrivent à une unité doivent être tels que cette unité se retrouve dans l'état que l'on veut mémoriser, même en l'absence de l'entrée extérieure qui a conduit à cet état, à condition que suffisamment d'autres unités du module soient déjà dans l'état à mémoriser. Autrement dit, on veut que le module puisse suppléer par son organisation interne à l'absence d'une partie de l'information externe.
Concrètement, cela donne la règle de modification suivante chez McClelland et Rumelhart :

$$\Delta w_{ij} = \mu \, \delta_i \, a_j$$

où $\mu$ est une constante qui règle la vitesse d'apprentissage, et $\delta_i$ est la différence entre l'entrée externe et l'entrée interne : $\delta_i = e_i - i_i$ .
Là encore, d'autres lois d'apprentissage ont été proposées, qui respectent le même objectif qualitatif que nous avons énoncé ci-dessus : nous aurons l'occasion de revenir dans la discussion sur la plus ou moins grande vraisemblance physiologique de ces règles.
Une simulation comporte alors les deux étapes suivantes :
- une période d'apprentissage, pendant laquelle on présente, un certain nombre de fois, une série de patterns d'entrée (appelée échantillon d'apprentissage), en modifiant après chaque stabilisation les poids du réseau.
- une période de test, où l'on présente de nouveaux patterns d'entrée, qui comportent des caractéristiques communes avec certains éléments de l'échantillon d'apprentissage : on mesure alors le succès de la mémorisation en comparant les réponses du réseau à ses réponses lors de l'apprentissage : si le même état stable est obtenu pour des patterns qui présentent des caractéristiques communes avec un pattern donné, on considérera que ce pattern a bien été mémorisé, et que les patterns utilisés lors du test ont conduit au rappel de cette forme « stockée » en mémoire.

## 1.3. Codage

L'un des aspects les plus importants du modèle de McClelland et Rumelhart concerne le type de codage qu'ils ont utilisé : il s'agit d'un codage « distribué », par opposition aux codages « localistes » dans lesquels chaque unité est destinée à représenter à elle seule un et un seul des items présentés au système. Ici au contraire, chaque unité est utilisée dans le codage de chaque item : ce n'est qu'en comparant les patterns globaux d'activité que l'on peut distinguer les différents items appris. Il n'y a donc pas l'équivalent de ce que l'on appelle « le neurone grand-mère », c'est-à-dire un neurone dont l'excitation produirait à elle seule l'évocation de sa grand-mère, et qui réciproquement ne serait excité que lors de cette évocation. C'est cette caractéristique du modèle qui lui a valu le nom, en intelligence artificielle, de système « subsymbolique » : les unités ne sont pas l'équivalent des symboles par lesquels on représente en intelligence artificielle classique les objets du monde ou les concepts, elles sont de niveau inférieur à ce niveau des symboles, dont l'équivalent est ici constitué par des combinaisons des valeurs des unités.
Concrètement, si le module comporte n unités (de l'ordre d'une vingtaine dans les simulations, comme nous l'avons dit, mais beaucoup plus élevé dans un modèle plus réaliste), un pattern d'entrée $\{e_i\}$ ou un pattern d'activité $\{a_i\}$ sont des vecteurs d'un espace à n dimensions. Chaque dimension peut être considérée comme une micro-caractéristique des items présentés. Ainsi si deux items se ressemblent beaucoup (i.e. ont beaucoup de micro-caractéristiques communes), les vecteurs correspondants seront proches, au sens de la distance euclidienne classique, dans l'espace des patterns d'entrée. Il y a donc une représentation topologique, en termes de proximité, de la notion de similarité entre les items présentés. C'est cette qualité qui explique les capacités d'adaptation du système : en gros le système va apprendre à répondre de la même manière à des entrées suffisamment proches. En fait, dans les simulations, les patterns significatifs en entrée sont des vecteurs binaires, dont les composantes valent $\pm 1$. Et les patterns d'activité (c'est-à-dire les réponses du système) sont composés de valeurs plus graduées, des valeurs franchement positives ou négatives (disons supérieures à 0,3 en valeur absolue) étant déclarées significatives, alors que des valeurs plus proches de 0 indiquent qu'il n'y a pas eu de « rappel » d'une forme apprise, du moins pour la composante en question.
Dans la plupart des expériences de McClelland et Rumelhart, l'espace d'entrée est divisé en deux parties, qui correspondent à des caractéristiques de nature différente des objets de l'environnement dont

les entrées sont théoriquement issues. L'exemple que donnent les auteurs, dont ils soulignent eux-mêmes le caractère simpliste, est celui d'entrées correspondant à des animaux, une partie des composantes du vecteur d'entrée servant à coder la forme visuelle de l'animal, tandis que l'autre servirait à coder son nom (ou toute autre étiquette associée à l'animal). Ainsi, l'une des expériences consiste à présenter au système des formes visuelles voisines (l'image d'un certain nombre de chiens), toutes associées à la même étiquette (le mot « chien »). Après apprentissage, si l'on présente la partie correspondant à l'étiquette (donc le mot « chien »), les autres composantes du vecteur d'entrée étant mises à 0, le pattern d'activité que l'on obtient sur les unités correspondantes consiste en une sorte de moyenne des formes visuelles apprises (le système a mémorisé une forme visuelle prototypique du chien, qui est rappelée quand on présente le mot « chien »).

## 2 Résultats

### 2. 1 Prototypes

Le premier type de résultats obtenus par McClelland et Rumelhart concerne en effet la capacité de ce système de mémoriser des formes prototypiques. Plus précisément, l'expérience consiste à se donner un vecteur binaire, dont les composantes sont choisies arbitrairement, qui servira de prototype, et d'engendrer à partir de ce vecteur une série de vecteurs voisins obtenus en changeant aléatoirement le signe de quelques unes des composantes du vecteur initial. On présente alors la série de vecteurs voisins comme échantillon d'apprentissage, et l'on constate que la forme qui a été mémorisée par le système (c'est-à-dire le vecteur d'activité qui est rappelé quand on présente une nouvelle entrée voisine, ou encore l'attracteur de la dynamique) est similaire au vecteur prototype, plutôt qu'aux exemples qui ont été présentés. Comme nous l'avons dit plus haut, si tous les exemples comportent une partie fixe (une étiquette commune), il suffit de présenter cette étiquette (les autres composantes étant nulles) pour que le prototype soit rappelé en entier. Inversement, si l'on associe à chaque exemple une étiquette différente (chaque chien est appelé par son nom propre, plutôt que par le nom générique « chien »), alors seule la partie du prototype ne correspondant pas à l'étiquette est rappelée (la forme visuelle du chien prototypique a été apprise, mais aucun nom ne lui est associé).
En fait, ce qui rend ces résultats non triviaux, c'est la capacité du système d'opérer sur plusieurs formes prototypiques simultanément. On fabrique un échantillon d'apprentissage qui mélange des vecteurs de trois familles, obtenues chacune par des modifications aléatoires d'un prototype différent (dans l'illustration des auteurs, il s'agit de chiens, de chats et de *bagels* — disons de petits gâteaux). Deux des familles (les chiens et les chats) possèdent de plus un certain nombre de caractéristiques communes, et donc se ressemblent plus qu'ils ne ressemblent à la troisième (les gâteaux). Le système s'avère alors capable de mémoriser les trois formes prototypiques : si l'on présente un vecteur proche de l'un quelconque des prototypes, c'est ce prototype qui constitue la réponse du système. Bien sûr si l'on ne présente que la partie commune aux deux familles proches (les chiens et les chats), les autres composantes étant à 0, on obtient un mélange des deux prototypes correspondants (une sorte d'intermédiaire entre chien et chat). Mais dès que l'on ajoute à l'entrée une petite partie des composantes capables de les discriminer, la réponse devient clairement le prototype le plus proche. La dynamique interne du système, obtenue par l'apprentissage, possède donc trois attracteurs correspondant aux trois prototypes.
Ainsi la preuve est faite que ce système est capable de « stocker » simultanément, dans un même ensemble de valeurs de poids, des informations différentes. Ceci est loin d'être intuitif, et conforte grandement l'hypothèse d'une mémoire diffuse, dans laquelle il est impossible de localiser l'information mémorisée : chaque poids de liaison, entre chaque couple d'unités, participe au stockage de l'information en mémoire, sans qu'il soit possible de déduire de la valeur d'un poids donné le rôle qu'il joue dans la mémorisation de telle ou telle information. En fait seule la configuration globale des poids est caractéristique de l'ensemble des informations stockées.

### 2.2 Mémoire spécifique

Il ne faudrait pas déduire de ce qui précède que ce système ne peut que mémoriser des informations

prototypiques, se comportant ainsi comme un système statistique uniquement capable d'extraire des informations moyennées à partir d'exemples, ce qui ne correspondrait pas bien sûr aux qualités de la mémoire humaine. En fait McClelland et Rumelhart ont aussi testé la capacité du modèle à retenir des exemples spécifiques. Pour cela, on constitue un échantillon d'apprentissage constitué pour partie d'une famille de vecteurs voisins obtenus comme précédemment (une série de chiens rencontrés au hasard des promenades...) et pour partie de deux vecteurs précis, voisins aussi du prototype, mais présents chacun en un grand nombre d'exemplaires dans l'échantillon d'apprentissage (le chien de la maison et le chien du voisin). Les résultats obtenus montrent que le système mémorise alors simultanément le prototype et les deux exemples fréquemment répétés : si l'on présente une entrée qui est plus proche du prototype que des deux exemples privilégiés, c'est le prototype qui est rappelé : mais dans le cas contraire, c'est l'exemple privilégié, le plus proche qui constitue la réponse du système. Du point de vue mathématique, cela signifie que la dynamique du système peut contenir plusieurs attracteurs très proches les uns des autres.
D'autres types de mesure sur le modèle permettent de mettre en évidence l'importance des exemples spécifiques. Il s'agit de prendre en compte les caractéristiques dynamiques du processus de rappel : en effet, un rappel peut être plus ou moins « fort », selon que les valeurs d'activité obtenues à la stabilisation sont plus ou moins grandes en valeur absolue ; et d'autre part le rappel peut être plus ou moins « rapide », selon que le processus de stabilisation réclame un plus ou moins grand nombre d'unités de temps. Grâce à ces mesures, on peut –sous certaines hypothèses que nous ne détaillerons pas ici– comparer le comportement du modèle avec celui de sujets humains dans des expériences de discrimination perceptive liée à des phénomènes de mémoire. On observe alors un effet de facilitation spécifique identique pour le modèle et pour l'humain quand l'item à discriminer a été présenté antérieurement (effet de « priming »), effet que l'on peut distinguer de l'effet dit de « familiarité » (qui facilite la reconnaissance d'items connus, comme les mots par rapport aux pseudo-mots), qui est lui aussi observable tant sur le modèle que chez l'humain. Quelles que soient les réserves que l'on peut émettre à propos de ces comparaisons, et sur lesquelles nous reviendrons, cela montre en tout cas que le modèle n'est pas réductible à un simple extracteur statistique de formes prototypiques.

## 2.3 Dysfonctionnements

Un autre aspect important du travail de McClelland et Rumelhart (1986b) concerne l'utilisation de ce modèle pour expliquer les comportements observables dans certains types d'amnésie. L'idée de base est d'étudier le comportement du modèle quand on modifie l'un de ses paramètres essentiels : la vitesse d'apprentissage $\mu$. On s'aperçoit alors qu'une réduction sensible de ce paramètre produit deux effets distincts : la capacité de mémoriser un exemple spécifique est pratiquement perdue, alors que la possibilité de mémoriser les prototypes demeure, même si cela devient plus long et plus difficile. Les auteurs font le rapprochement avec les cas d'amnésie où le patient est incapable de se souvenir d'événements spécifiques, alors qu'il peut encore « apprendre », ou plutôt acquérir des savoir-faire, dans certains domaines particuliers : ainsi le célèbre patient H.M., pourtant atteint d'une amnésie très sévère, a-t-il été capable d'apprendre à résoudre le problème des Tours de Hanoï.
Bien entendu les auteurs reconnaissent que le modèle très simple qu'ils ont construit ne saurait à lui seul expliquer toutes les caractéristiques de l'amnésie. En particulier, deux types de comportements observés semblent à première vue contradictoires avec les principes qui ont présidé à la conception du modèle. Il s'agit d'une part des cas d'amnésie où les souvenirs les plus anciens sont conservés alors que le souvenir des événements récents est impossible, et d'autre part des cas de rétablissement où ces événements récents, qui semblaient non mémorisés, redeviennent accessibles à la mémoire consciente. Il est clair que dans le modèle présenté ci-dessus de tels phénomènes sont inexplicables. Pour tenter de remédier à ces difficultés, les auteurs proposent un modèle plus complexe, obéissant aux mêmes principes de base (en particulier au « stockage distribué » de la mémoire dans les poids du réseau), mais dans lequel le mécanisme de modification des poids dépend de plusieurs paramètres. L'un d'entre eux représente un « modulateur », dont la présence serait nécessaire pour que les modifications des poids soient « consolidées » et les modifications non consolidées déclinent rapidement avec le temps. Ils montrent alors que certaines amnésies seraient modélisables par une baisse de ce paramètre de

consolidation, et qu'en particulier les phénomènes évoqués plus haut auraient alors leurs correspondants dans le modèle.
Dans le même ordre d'idées, Horn *et al.* (1993) ont utilisé des données neurophysiologiques sur la maladie d'Alzheimer, pour construire un modèle connexionniste (du type réseau de Hopfield), dans lequel on simule deux caractéristiques anatomiques observées dans cette maladie : une diminution du nombre de synapses par neurone, et une relative invariance de la surface synaptique, qu'ils interprètent comme un phénomène de compensation, les synapses restantes augmentant leur surface de contact. Ils montrent alors que le même modèle peut se comporter de deux façons différentes, suivant les relations qu'entretiennent le phénomène de destruction de synapses et le mécanisme de compensation. Dans un cas, on assiste à une chute brutale des performances du réseau après une période plus ou moins longue de préservation des capacités de mémorisation, alors que dans l'autre cas, on observe une dégradation progressive des performances. Ces auteurs font remarquer que ces résultats sont conformes aux observations cliniques : chez des patients jeunes ou au contraire très âgés, on a décrit un progrès très rapide de la maladie, tandis qu'elle présente une évolution beaucoup plus graduelle chez la majorité des patients. Ces différents cas correspondraient aux différentes relations possibles entre destruction synaptique et mécanisme compensatoire : fortes capacités compensatoires chez les jeunes, absence totale de compensation chez les très âgés, compensation variable, augmentant avec le pourcentage de synapses détruites, dans le cas le plus fréquent.

## 3. Discussion

### 3.1. Capacité des réseaux

On peut d'abord se demander si le système, au delà de la maquette illustrative présentée (qui ne comporte, rappelons-le, qu'une vingtaine d'unités et qui n'est utilisée que pour mémoriser trois ou quatre formes), possède une capacité suffisante, à la fois en quantité et en qualité, pour dépasser le stade de ces exemples « jouets ».
En ce qui concerne la quantité, des travaux théoriques (Amit, 1989) ont permis de déterminer précisément le nombre maximum de formes que l'on peut stocker en mémoire dans un tel réseau : plus exactement sur les réseaux dits « de Hopfield » (Hopfield, 1982, 1984), qui sont assez proches des réseaux présentés ici. Si l'on appelle N le nombre d'unités d'un réseau de Hopfield, le nombre maximum de formes mémorisables est de l'ordre de 0,14 N. Ces résultats sont très encourageants. Si l'on pense que le modèle réaliste envisagé par McClelland et Rumelhart pourrait comporter jusqu'à des millions d'unités, cela signifie que des centaines de milliers d'informations différentes seraient mémorisables, ce qui laisse une marge confortable sur l'ordre de grandeur de la quantité d'information mémorisable souhaitée. Il faut noter d'ailleurs que l'on peut obtenir des capacités de mémorisation beaucoup plus importantes avec des architectures voisines (Gardner, 1988), ou même, avec la même architecture, mais des hypothèses de fonctionnement légèrement différentes (Tsodyks et Feigel'Man, 1988 ; Buhmann *et al.*, 1988). Ainsi, le résultat que nous venons de donner (0,14 N) est obtenu en supposant que les valeurs de +1 et –1 sont équiprobables dans les vecteurs à mémoriser ; mais si l'on suppose qu'il y a au contraire une profonde dissymétrie (par exemple, on suppose qu'il y a en moyenne beaucoup moins de +1 que de –1, ce qui correspondrait au fait que chaque forme à mémoriser n'excite qu'un petit nombre d'unités), on obtient alors un nombre maximum de formes mémorisables qui peut largement dépasser N.
Du point de vue qualitatif par contre, les auteurs eux-mêmes mettent en évidence une limite du modèle : les patterns qui peuvent être appris par le réseau ne sont pas quelconques : ils doivent obéir à une contrainte, dite de « prédictabilité linéaire », qui limite sérieusement la puissance du système. Mais les auteurs suggèrent une architecture un peu plus complexe, qui consiste essentiellement en l'ajout dans le réseau « d'unités cachées » (en fait d'unités ne recevant pas d'entrée extérieure), et qui suffit à circonvenir à cette difficulté. Des travaux théoriques ont montré qu'en utilisant des règles d'apprentissage plus sophistiquées (voir par exemple Pineda, 1987), n'importe quelle configuration pouvait être apprise par un réseau comportant de telles unités cachées. Ainsi, là encore, il ne semble pas y avoir de problèmes insurmontables : du point de vue technique, les réseaux sont des outils tout à

fait adaptés à ce genre de tâches.
Reste que ces résultats sur la capacité des réseaux sont pour l'instant purement théoriques. En fait, les seuls réseaux (de ce type : c'est-à-dire entièrement interconnectés, par opposition aux réseaux unidirectionnels à couches, beaucoup plus utilisés et mieux maîtrisés) qui aient été implémentés n'ont qu'un nombre ridiculement petit d'unités (au plus quelques centaines). Et encore, l'application des algorithmes d'apprentissage, en particulier en présence d'unité cachées, pose des problèmes non encore complètement résolus. Or on sait que les questions de taille en informatique sont loin d'être innocentes : bien des systèmes, fonctionnant correctement sous forme de maquettes, se révèlent inutilisables quand on passe à des applications « en vrai grandeur ». Cette situation est en partie due à des questions de matériel : il est vraisemblable que l'arrivée sur le marché de machines « massivement parallèles » va permettre un bond en avant dans la maîtrise des réseaux de grande taille. Mais quoi qu'il en soit, on peut rester sceptique quant aux chances de réaliser dans les prochaines années des systèmes de la taille envisagée ici.

### 3.2. Vraisemblance du modèle

Au plan neurophysiologique, les « neurones formels » du réseau sont évidemment des représentations caricaturales, par leur simplisme, des neurones réels, que ce soit du point de vue de leur fonctionnement ou du point de vue de leur architecture. Mais en fait au-delà de cette simplification, le problème est plutôt de se demander s'il y a dans le modèle des aspects qui sont franchement contradictoires avec les connaissances de la neurophysiologie : après tout, pour le reste, si un modèle simpliste est déjà capable de comportements intéressants, la mise en place d'unités plus complexes ne peut qu'augmenter les capacités du système. Si l'on s'en tient donc à rechercher ce qui pourrait constituer des contradictions rédhibitoires, deux points méritent d'être examinés. D'abord, les unités du modèle peuvent posséder des poids de connexion positifs vers certains de leurs voisins et inhibiteurs vers d'autres, alors que l'on sait que toutes les synapses provenant d'un même neurone sont de même nature (excitateurs ou inhibiteurs). Mais cette anomalie peut être –plus ou moins facilement– corrigée dans le modèle : sans entrer dans les détails, l'idée est d'ajouter systématiquement l'équivalent d'interneurones inhibiteurs pour gérer les différences de signe des poids. L'autre point concerne l'apprentissage. La règle choisie par McClelland et Rumelhart est très peu plausible : le renforcement d'une synapse réclamerait le calcul de la différence entre les excitations internes et externes du neurone postsynaptique ! Mais, là aussi, d'autres règles d'apprentissage, moins extravagantes, peuvent remplacer celle-ci sans dommage : il est symptomatique par exemple que la simple règle de Hebb (1949), dont la réalité physiologique est depuis longtemps attestée, conduise à des résultats tout à fait honorables. Ainsi, en ce qui concerne la physiologie, bien que l'on soit encore loin d'une modélisation un tant soit peu réaliste de la complexité des phénomènes, il n'y a pas de problème majeur qui conduise à rejeter d'emblée la voie ouverte par ce type de modèle.
Il n'en est pas de même en ce qui concerne ce que l'on pourrait appeler la « vraisemblance psychologique » du modèle, et c'est sûrement sur ce plan que le modèle est le plus discutable. En effet, par sa conception même, même s'il ne fonctionne pas comme un simple extracteur de prototype, ce système est avant tout un système d'apprentissage statistique, plus qu'un modèle de mémoire. S'il est capable de mémoriser des exemples spécifiques aussi bien que des prototypes, ce n'est que dans la mesure où ces exemples spécifiques sont répétés un grand nombre de fois : rien à voir avec les capacités de la mémoire humaine, où une seule présentation peut suffire à inscrire définitivement un événement. On peut bien sûr rétorquer que le système complet (des centaines de modules en interaction) permettrait d'obtenir ce comportement, en arguant du fait que l'événement en question est alors mis en relation avec tout un ensemble d'autres informations déjà mémorisées. Mais cela reste hypothétique, et rien ne dit que les mécanismes modélisés ici seront alors suffisants, ni même essentiels : ils ne joueront peut-être qu'un rôle secondaire par rapport à d'autres processus qu'il faudra mettre en place pour pouvoir obtenir le comportement souhaité.
Il paraît en tout cas très prématuré de commencer à comparer les performances du système avec des données psychologiques ou psychopathologiques. Que peuvent donc bien signifier des comparaisons de vitesse ou de force de rappel ou –pire encore– de mettre en relation une modification de paramètre

avec des comportements d'amnésie, alors que l'on annonce dès le départ que ce modèle représente, au mieux, un système partiel (et considérablement réduit en taille) qui n'a de sens que s'il est plongé dans un système de niveau supérieur dont il ne constitue qu'un élément parmi des milliers ? Il y a là un « raccourci » qui enlève toute crédibilité à cet aspect précis du travail de modélisation dont on peut douter de l'intérêt scientifique.

### 3.3. « Epistémologie expérimentale »

Est-ce à dire que ce type de modélisation ne présente aucun intérêt tant qu'on ne pourra pas présenter de système plus réaliste ? Loin de là : on peut même affirmer que ces petites maquettes sont sûrement plus intéressantes par elles-mêmes que ne le serait un système plus complet, construit de toutes pièces, qui aurait peu de chances de toutes façons de correspondre à la réalité physiologique. En effet, l'intérêt de ces modèles ne réside pas dans leur capacité à rendre compte de performances quantitatives mesurables. Il est avant tout de tester des idées, d'ouvrir des pistes de réflexion qui peuvent être utiles aux spécialistes des disciplines concernées. Ainsi, le fait d'avoir pu exhiber un système dans lequel quelque chose qui ressemble à de la mémoire n'est pas matérialisé dans des unités dédiées chacune à une information donnée, mais distribué dans des poids de connexion d'un ensemble complet d'unités, sans que l'on puisse détecter où se situe dans cet ensemble chacune des informations mémorisées, voilà qui donne effectivement à réfléchir, et qui peut aider psychologues et physiologistes à imaginer des expériences permettant de tester si un tel principe organisationnel est à l'oeuvre ou non dans la mémoire humaine. De même, si le modèle de la maladie d'Alzheimer présenté plus haut ne saurait être considéré comme réaliste, son intérêt n'en est pas moins grand. En effet, il prouve qu'un même processus (destruction synaptique et mécanisme compensatoire) peut produire des effets très différents (chute brutale ou déclin progressif des capacités de mémorisation). Cela indique donc au neuropsychologue qu'il n'est pas obligé de rechercher des explications différentes pour rendre compte des diverses formes que peut prendre cette maladie.

Un dernier exemple, qui n'est pas directement axé sur le problème de la mémoire puisqu'il s'agit d'un modèle de la dyslexie profonde, mais dont l'importance dépasse largement ce sujet précis : Hinton et Shallice (1991) ont construit un réseau connexionniste, d'une architecture assez différente de celles que nous avons présentées ici (mais fonctionnant aussi comme une dynamique à attracteurs), censé simuler le processus de lecture : les patterns d'entrée correspondent à des mots (plus précisément la forme orthographiée de ces mots), et les patterns de sortie sont appariés à des ensembles de traits sémantiques (le « sens » de ces mots). Ils montrent alors qu'en détériorant une partie quelconque du réseau, que ce soit en amont (dans la partie directement reliée à « l'input visuel ») ou en aval (dans les centres de traitement « sémantique »), on obtient un comportement identique, qui ressemble à celui d'un grand nombre de ces dyslexiques profonds. Des erreurs de type « sémantique » (par exemple, le mot présenté est *cat* et le mot « compris » est *mice*) voisinent avec des erreurs de type « visuel » (le mot présenté est *mat* et le mot compris est *cat*), avec une proportion très élevée d'erreurs « mixtes » (le mot présenté est *rat* et le mot compris est *cat*). Bien que, là encore, le modèle ne soit absolument pas réaliste, ce résultat est très important, parce qu'il prouve que la localisation de la lésion (plus ou moins périphérique) n'a aucune incidence sur le type d'erreurs observées : voilà qui remet en question bien des hypothèses de bon sens couramment admises. On aurait pu penser que des erreurs de nature différente étaient forcément dues à des lésions localisées dans des régions fonctionnelles différentes, et réciproquement que des détériorations impliquant des niveaux différents du processus provoqueraient des erreurs de nature différente (une lésion périphérique provoquant plutôt des erreurs « visuelles », et une lésion plus centrale des erreurs « sémantiques »).

Autrement dit, l'intérêt de ces modèles ne tient pas à leur capacité de simuler de façon plausible tel ou tel aspect du comportement cognitif humain, mais plutôt à proposer des hypothèses nouvelles, concrétisées par des maquettes qui montrent la fécondité de certaines structures fonctionnelles, qui peuvent parfois paraître contre-intuitives. A ce titre, ces maquettes doivent rester simples, pour ne pas masquer dans une foule de paramètres incontrôlables les principes de base qui expliquent leurs performances. Il nous semble que si ces objectifs sont clairs, cette approche, relativement originale dans les sciences d'aujourd'hui, que l'on pourrait appeler « épistémologie expérimentale », peut se

révéler extrêmement fructueuse : c'est l'un des intérêts principaux du connexionnisme que d'avoir introduit, par l'intermédiaire de travaux comme ceux de McClelland et Rumelhart, cette nouvelle façon d'aborder les problèmes de la cognition.

## Bibliographie


AMIT D.J. (1989), *Modeling brain function : the world of attractor neural networks*, Cambridge University Press.

BUHMANN J., DIVKO R., SCHULTEN K. (1988), Associative memory with high information content, *Physical Review*, A39, 2689.

HEBB D.O. (1949), *The organization of behavior*, New York, Wiley.

HINTON G.E., SHALLICE T. (1991), Lesioning an attractor network: investigations of acquired dyslexia, *Psychological Review*, 98, 1.

GARDNER E. (1988), The space of interactions in neural network models, *Journal of Physics*, 21A, 257.

HOPFIELD J.J. (1982), Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences*, USA, 79, 2554-2558.

HOPFIELD J.J. (1984), Neurons with graded response have collective computational properties like those of two-states neurons, *Proceedings of the National Academy of Sciences*, USA, 81, 3088-3092.

HORN D., RUPPIN E., USHER M., HERRMANN M. (1993), Neural network modeling of memory deterioration in Alzheimer's disease, *Neural Computation*, 5, 736-749.

MCCLELLAND J.L., RUMELHART. D.E. (1986a), A distributed model of human leaming and memory, in J.L. McClelland, D.E. Rumelhart & the PDP Research Group (Eds), *Parallel Distributed Processing*, Cambridge, Mass., MIT Press, 2, 170-215.

MCCLELLAND J.L., RUMELHART. D.E. (1986b), Amnesia and distributed Memory, in J.L. McClelland, D.E. Rumelhart & the PDP Research Group (Eds), *Parallel Distributed Processing*, Cambridge, Mass., MIT Press, 2, 503-528.

PINEDA F.J. (1987), Generalization of back propagation to recurrent neural networks, *Physical Review Letters*, 59, 2229-2232.

SHALLICE T. (1991), Précis of From neuropsychology to mental structure, *Behavioral and Brain Sciences*, 14, 429-469.

TSODYKS M.V., FEIGEL'MAN M.V. (1988), The enhanced storage capacity in neural networks with low activity level, *Europhysical Letters*, 46, 101-105.