UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'ARCHITECTURE COGNITIVE MODULAIRE ET INTERACTIVE SOUS LE PRINCIPE D'ÉNERGIE LIBRE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

MAITRISE EN PHILOSOPHIE

PAR

SIMON TREMBLAY

JUILLET 2022

REMERCIEMENTS

Je souhaite remercier les nombreuses personnes qui ont contribué d'une façon ou d'une autre à l'avancement de ce projet de recherche et qui lui ont permis d'aboutir. En particulier, je remercie mon directeur, Pierre Poirier, non seulement pour ses qualités de directeur, mais aussi pour ses qualités humaines. Surtout, je remercie Safae, mon éternelle complice, sans qui rien de tout cela n'aurait été possible : « *Sed omnia præclara tam difficilia quam rara funt* ». Finalement, merci au sourire d'Alaric, qui illumine chacune de mes journées.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

FEP	*Free Energy Principe* (Principe d'énergie libre)
HMM	*Hierarchically Mechanistic Mind* (Esprit hiérarchiquement mécaniste)
MMH	*Massive Modularity Hypothesis* (Hypothèse de la modularité massive)
MRH	*Massive Redeployment Hypothesis* (Hypothèse du redéploiement massif)
EST	*Evolutionary System Theory* (Théorie des systèmes évolutionnaires)
LOT	*Language Of Thought* (Langage de la pensée)
NRP	*Neuroscientifically Relevant Psychological factors* (Facteurs psychologiques neuroscientifiquement pertinents)
TALoNs	*Transiently Assembled Local Neuronal Subsystems* (Assemblages transitoires de sous-systèmes neuronaux locaux)
IDS	*Interactive Differentiation and Search* (Différenciation et recherche interactive)

RÉSUMÉ

Dans ce mémoire, nous proposons d'envisager les thèses principales de l'hypothèse de la modularité massive et de l'hypothèse du redéploiement massif, comme étant ultimement complémentaire dans la perspective d'un pluralisme explicatif intégratif sous le principe d'énergie libre. D'une part, l'hypothèse de la modularité massive est une proposition issue du paradigme cognitiviste. Elle soutient que l'architecture cognitive repose sur des modules fonctionnellement spécialisés qui ont été façonné par des pressions sélectives pour la résolution de problèmes adaptatifs dans l'environnement évolutionnaire. D'autre part, l'hypothèse du redéploiement massif est une proposition issue du paradigme de la cognition incarnée. Elle soutient que l'architecture cognitive repose sur des biais corticaux fonctionnellement différenciés qui ont été façonné par des pressions sélectives pour le redéploiement des ressources neuronales dans un environnement largement conservé entre les générations. Nous soutiendrons que le principe d'énergie libre permet d'envisager un pluralisme explicatif intégratif qui repose sur les niveaux d'analyse biologique proposé par Tinbergen. Dans le cadre de ce pluralisme explicatif intégratif, nous distinguons entre les niveaux d'analyse associés à l'explication fonctionnaliste sélectionniste, qui procède des « parties au tout », et les niveaux d'analyse associés à l'explication structuraliste développementaliste, qui procède du « tout aux parties ». Ce faisant, nous proposons d'associer l'hypothèse de la modularité massive au fonctionnalisme en raison de son emphase sur la sélection naturelle et l'hypothèse du redéploiement massif au structuralisme en raison de son emphase sur l'auto-organisation. Puisque les niveaux d'analyse biologique concernent les échelles temporelles sur lesquelles la sélection naturelle et l'auto-organisation interagisssent, nous soutiendrons que sur une échelle temporelle évolutionnaire, les pressions sélectives locales fasçonnent des modules spécialisés aux niveaux hiérarchiques inférieurs, et que sur une échelle temporelle intergénérationnelle, les pressions sélectives globales fasçonnent des hubs connecteurs « rich-club » fonctionnellement différenciés aux niveaux hiérarchiques supérieurs.

Mots clés : modularité massive, redéploiement massif, cognitivisme, cognition incarnée, énactivisme, principe d'énergie libre, pluralisme explicatif

INTRODUCTION

Depuis quelques années, nous assistons à ce que certains voient comme un changement de paradigme en sciences cognitives : plusieurs des thèses centrales de l'orthodoxie cognitiviste sont contestées par l'émergence du paradigme de la cognition incarnée. D'une part, le paradigme cognitiviste soutient que le cerveau est un système computationnel de traitement de l'information. Ce paradigme implique une conception internaliste de la cognition, supposant que celle-ci se limite au système nerveux. Par conséquent, elle serait essentiellement un processus qui dépend de la manipulation de représentations découplées de l'environnement. D'autre part, l'approche incarnée de la cognition soutient que le cerveau est un système dynamique de contrôle. Ce paradigme implique une conception externaliste de la cognition, supposant que celle-ci s'étend du cerveau jusqu'au corps et même à l'environnement. Par conséquent, elle serait un processus qui repose sur des interactions dynamiquement couplées avec l'environnement. Ce mémoire propose de dépasser l'opposition entre le cognitivisme classique et approches incarnées de la cognition en faisant appel au principe d'énergie libre (FEP) de Karl Friston (2010).

Le FEP est un énoncé formel qui stipule que les systèmes biologiques parviennent à maintenir leur organisation en minimisant une quantité informationnelle nommée énergie libre. Plutôt toutefois que d'essayer de simplement assimiler le FEP à l'un ou l'autre de ces deux paradigmes, comme l'ont fait d'autres chercheurs (Hohwy, 2017; Kiefer & Hohwy, 2018; Bruineberg et al., 2018; Ramstead & Kirchhoff, 2020), nous soutiendrons qu'il serait plus avantageux d'interpréter le FEP dans la perspective d'un pluralisme explicatif intégratif qui envisage la complémentarité des paradigmes en question. Pour concrétiser cette proposition, nous étudierons deux hypothèses architecturales opposées dans la littérature, l'une représentant le cognitivisme – soit l'hypothèse de la modularité massive[1], ou Massive Modularity Hypothesis (MMH), (Carruthers, 2006; 2008) – et l'autre représentant l'approche incarnée – soit l'hypothèse du redéploiement massif[2], ou Massive Redeployment Hypothesis (MRH), (Anderson, 2014; 2016). Nous soutiendrons que, comprises à la lumière du FEP, ces hypothèses architecturales ne

[1] L'hypothèse de la modularité massive est une conception cognitiviste qui suppose que la cognition est un processus qui repose sur le système nerveux. En effet, il s'agit d'une conception reconstructive de la cognition comprise comme la manipulation de représentations qui sont découplées de l'environnement.

[2] L'hypothèse du redéploiement massif est une conception incarnée de la cognition qui suppose que la cognition est un processus qui repose sur le système cerveau-corps-environnement. En effet, on propose une conception performative de la cognition comprise comme la gestion des interactions qui sont couplées avec l'environnement.

s'opposent pas : bien au contraire, elles rendent en fait toutes deux compte d'aspects interdépendants de l'architecture cognitive. Selon la perspective cognitiviste de la MMH, l'architecture fonctionnellement spécialisée repose principalement sur des modules « domaine-spécifiques » – c'est-à-dire des mécanismes chargés d'accomplir des fonctions particulières et non plusieurs fonctions à la fois – résultant de pressions sélectives pour la résolution de problèmes adaptatifs dans l'environnement évolutionnaire (Carruthers, 2006; 2008). Quant à la perspective radicalement incarnée de la MRH, l'architecture fonctionnellement différenciée repose sur le redéploiement des mêmes propriétés fonctionnelles « domaine-générales » – c'est-à-dire des mécanismes chargés d'accomplir plusieurs fonctions à la fois et non des fonctions particulières – résultant de pressions sélectives pour le redéploiement de ressources neuronales dans l'environnement largement conservé entre les générations[3] (Anderson, 2014; 2016).

Dans ce mémoire, nous soutiendrons l'idée selon laquelle le FEP se trouve au carrefour des approches cognitivistes et radicalement incarnées de la cognition en envisageant les thèses qui caractérisent la MMH et la MRH comme étant complémentaires dans un pluralisme explicatif intégratif[4] qui repose sur les niveaux d'analyse de Tinbergen (1963). Selon la proposition avancée dans ce mémoire, l'architecture cognitive peut être expliquée en fonction de deux stratégies explicatives distinctes, mais ultimement complémentaires : (1) la stratégie du fonctionnalisme sélectionniste, qui procède des « parties au tout » et (2) la stratégie du structuralisme développementaliste, qui procède du « tout aux parties » (voir Witherington et Lickliter, 2016). En d'autres termes, nous distinguons entre constitution « bottom-up », qui repose sur des interactions causales-mécanistes, et contrainte « top-down », qui repose sur des

[3] Pour Anderson (Anderson & Finlay, 2014), il n'est pas nécessaire de postuler une évolution mosaïque qui permet aux pressions sélectives de cibler des régions individuelles. L'évolution du cerveau se ferait de façon concertée, les pressions sélectives sur les processus développementaux pour permettre un large éventail de propriétés fonctionnelles pouvant être adaptativement redéployées dans diverses configurations fonctionnelles. Les biais fonctionnels « typiques à l'espèce » sont héritables par transmission épigénétique/exogénétique intergénérationnelle de l'information en collaboration avec des afférences sensorielles hautement stéréotypées et des expériences initiales largement similaires dans l'environnement largement conservé entre les générations.

[4] Nous soutenons que le FEP s'accorde avec un pluralisme explicatif intégratif qui distingue entre les différents niveaux d'analyse de Tinbergen (1963). Il s'agit d'un pluralisme explicatif intégratif puisque les niveaux d'analyse concernent des intérêts explicatifs divergents, mais ultimement complémentaires pour une explication complète des phénomènes biologiques (voir Mitchell, 1992). Plus précisément, nous proposons une distinction supplémentaire entre deux grandes stratégies explicatives : (1) les explications fonctionnalistes et (2) les explications structuralistes. Le FEP permet l'intégration des stratégies explicatives puisque que son formalisme mathématique explique les propriétés structurelles émergentes des systèmes dynamiques autonomes, mais fournit aussi des contraintes énergétiques sur l'organisation causale des mécanismes biologiques qui les constituent. Puisqu'il existe plusieurs façons de réaliser les mécanismes en fonctions de ces contraintes énergétiques, les deux stratégies fournissent deux manières distinctes, mais complémentaires d'expliquer un même phénomène cible – c'est-à-dire les processus biologiques complexes qui sous-tendent la cognition.

propriétés systémiques émergentes (qui sont non-causales au sens traditionnel de causalité mécaniste ou efficiente) (voir Moreno & Suarez, 2020). En effet, sans référence à une notion de contrainte « top-down », il semble impossible d'expliquer comment certains états spécifiques sont réalisés parmi des ensembles d'états possibles de plus bas niveau et comment cette sélection de plus haut niveau génère des propriétés structurelles émergentes qui permettent l'intégration d'un tout autoorganisé qui est plus que la somme des parties. Comme nous le verrons plus en détail au chapitre 1, les stratégies explicatives fonctionnaliste (à ne pas confondre avec la théorie sur la nature des états mentaux en sciences cognitives) et structuraliste sont représentées dans plusieurs débats théoriques importants dans l'histoire de la biologie. D'une part, la stratégie fonctionnaliste repose sur une conception causale-mécaniste d'une matière biologique passive façonnée de façon extrinsèque par la sélection naturelle (Linde Medina, 2010; Witherington & Lickliter, 2016). Cette stratégie explicative est associée à la MMH puisqu'elle soutient que l'architecture cognitive est façonnée par des pressions sélectives locales sur des modules spécialisés dans la résolution des problèmes adaptatifs de l'environnement évolutionnaire. D'autre part, la stratégie structuraliste repose sur une conception organiciste d'une matière biologique active façonnée de façon intrinsèque par l'auto-organisation (Linde Medina, 2010; Witherington & Lickliter, 2016). Cette stratégie explicative est associée à la MRH puisqu'elle soutient que l'architecture cognitive est façonnée par des pressions sélectives globales qui agissent sur les processus développementaux pour permettre le redéploiement des ressources neuronales dans des assemblages flexibles qui s'adaptent rapidement aux changements environnementaux. Dans la perspective d'une théorie des systèmes évolutionnaires, ou Evolutionary Systems Theory (EST), nous soutiendrons que les niveaux d'analyse de Tinbergen correspondent aux échelles temporelles sur lesquelles la sélection naturelle et l'auto-organisation interagissent (Badcock, 2012; Badcock et al., 2019). Plus particulièrement, nous soutiendrons que la stratégie explicative du fonctionnalisme sélectionniste résulte d'une emphase sur les niveaux d'analyse du mécanisme (causation) et de la fonction (adaptation) et que la stratégie explicative du structuralisme développementaliste résulte d'une emphase sur les niveaux d'analyse de l'ontogénie (développement) et de la phylogénie (évolution)[5].

Tel que précédemment mentionné, la MMH et la MRH représentent deux conceptions traditionnellement opposées de l'architecture cognitive, l'une étant associée au paradigme cognitiviste et l'autre étant

[5] Dans la perspective de la théorie des systèmes évolutionnaires de Paul Badcock (Badcock et al., 2019), « [...] in this context, phylogeny is used to refer to the intergenerational processes responsible for producing evolutionary change within a species, not the evolutionary outcomes of such processes (e.g., our species' position on a phylogenetic tree) ».

associée au paradigme de la cognition incarnée. Leur opposition peut être décomposée en trois ensembles de thèses souvent considérées dans la littérature comme étant opposées (pour un survol des oppositions théoriques entre le cognitivisme et la cognition incarnée, voir Wilson & Foglia, 2011) : (1) l'opposition entre computationnalisme représentationnel et dynamicisme non-représentationnel, (2) l'opposition entre modularité quasi-décomposable et interactivité non-décomposable, et enfin (3) l'opposition entre l'adaptation par évolution mosaïque et l'adaptativité par évolution concertée[6]. Comme nous le verrons au chapitre 1 quand nous aborderons chacune des thèses individuellement, le paradigme cognitiviste internaliste repose sur des composantes locales sélectionnées alors que le paradigme de la cognition incarnée repose sur des processus globaux autoorganisés. Dans le cadre d'un pluralisme explicatif intégratif, nous soutiendrons que les thèses complémentaires qui caractérisent la MMH peuvent être comprises dans la perspective d'une application d'un fonctionnalisme biologique à l'architecture cognitive. Dans l'ensemble, les thèses du computationnalisme représentationnel, de la modularité quasi-décomposable et de l'adaptation par évolution mosaïque concernent les composantes fonctionnelles locales dans l'architecture cognitive. En effet, la MMH soutient que l'architecture cognitive est *quasi-décomposable*, c'est-à-dire que les interactions dans les composantes sont plus importantes que les interactions entre les composantes (dominance componentielle; voir Van Orden, Holden et Turvey, 2003). Dans cette perspective fonctionnaliste, les fonctions globales sont construites par les fonctions des composantes locales (des parties au tout). Toujours dans le cadre du même pluralisme explicatif intégratif, nous soutiendrons que les thèses complémentaires qui caractérisent la MRH peuvent être interprétées dans la perspective d'une application d'un structuralisme biologique à l'architecture cognitive. Dans l'ensemble, les thèses du dynamicisme non-représentationnel, de l'interactivité non-décomposable et de l'adaptativité par évolution concertée concernent les processus fonctionnels globaux dans l'architecture. En effet, la MRH soutient que l'architecture cognitive est *non-décomposable*, c'est-à-dire que les interactions entre les composantes sont plus importantes que les interactions dans les composantes

[6] Nous distinguons ici entre l'adaptation, qui désigne un trait adapté à l'environnement ancestral, et l'adaptativité, qui désigne un trait qui s'adapte à l'environnement actuel. L'adaptation est associée au néo-darwinisme (de la psychologie évolutionniste, par exemple), qui repose essentiellement sur la sélection naturelle et la transmission génétique. Selon ce modèle, le cerveau évolue de façon mosaïque (ou modulaire) puisque les pressions sélectives ciblent des traits phénotypiques de façon indépendante (voir Barton & Harvey, 2000). L'adaptativité est associée à la synthèse évolutionnaire étendue (de la biologie évolutionnaire-développementale par exemple), qui repose essentiellement sur l'auto-organisation et la transmission épigénétique/exogénétique. Selon ce modèle, le cerveau évolue de façon concertée (ou non-modulaire) puisque les pressions sélectives ciblent les processus développementaux qui façonnent un tout intégré (voir, Finlay & Darlington, 1995; Anderson & Finlay, 2014). Évidemment, on peut comprendre l'adaptativité comme une forme d'adaptation puisqu'elle est régulée/maintenue par des pressions sélectives (en excluant les formes de plasticité non-adaptatives).

(dominance interactive; voir Van Orden, Holden et Turvey, 2003). Dans cette perspective structuraliste, ce sont les fonctions globales qui contraignent les fonctions des composantes locales (du tout aux parties).

Comme nous le verrons au chapitre 2 lorsque nous aborderons l'architecture cognitive dans la perspective d'un pluralisme explicatif intégratif sous le FEP (Badcock et al. 2019), le cerveau est compris comme un système hiérarchique adaptatif complexe qui minimise l'énergie libre variationnelle[7] par le biais de cycles action-perception. Selon cette hypothèse, les cycles action-perception sont générés par des interactions dynamiques bidirectionnelles entre mécanismes neurocognitifs « domaine-spécifiques » hautement ségrégués aux niveaux inférieurs de la hiérarchie ainsi que des mécanismes neurocognitifs « domaine-généraux » hautement intégrés aux niveaux supérieurs de la hiérarchie. Cette hypothèse de l'architecture cognitive, nommé l'esprit hiérarchiquement mécaniste ou Hierarchically Mechanistic Mind (HMM), intègre également la EST qui explique le phénotype comme émergeant de la minimisation de l'énergie libre sur différentes échelles temporelles (Badcock, 2012; Badcock et al., 2019). En effet, selon cette théorie, l'interaction entre les principes généraux de la sélection naturelle et de l'auto-organisation façonne le phénotype sur les quatres échelles temporelles qui correspondent aux niveaux d'analyse biologique de Tinbergen (1963), soit l'adaptation à l'échelle temporelle évolutionnaire, la phylogénie[8] à l'échelle temporelle intergénérationnelle, l'ontogénie à l'échelle temporelle développementale et le mécanisme en temps réel (Badcock, 2012; Badcock et al., 2019). En ce sens, la EST permet l'intégration hiérarchique des paradigmes explicatifs en psychologie qui se concentrent sur différents niveaux d'analyse biologique : les explications fonctionnelles pour les caractéristiques adaptatives « typiques de l'espèce » qu'on associe à la psychologie évolutionniste et la synthèse moderne néo-darwinienne; les explications intergénérationnelles des similitudes et différences entre les groupes qu'on associe ici à la psychologie évo-dévo et la synthèse évolutionnaire étendue; les explications du développement individuel qu'on associe à la psychologie développementale; et les explications mécanistes pour les phénomènes

[7] Un principe variationnel permet la résolution de problèmes d'optimisation sous contraintes, ce qui peut s'exprimer par le principe de temps moindre (optique) de Fermat, selon lequel « la nature agit toujours par les voies les plus courtes » ou encore par le principe d'action moindre (mécanique) d'Hamilton, selon lequel « la nature agit toujours par le chemin de moindre résistance ».

[8] Dans la perspective de la théorie des systèmes évolutionnaires de Paul Badcock (Badcock, 2012; Badcock et al., 2019), la phylogénie réfère au processus causal dynamique de transmission épigénétique/exogénétique de caractères entre les générations (plutôt qu'au positionnement dans l'arbre de la vie). Dès lors, on s'intéresse aux causes des changements phylogénétiques opérant sur une échelle temporelle intergénérationnelle.

neurocognitifs en temps réel qu'on associe aux diverses sous-disciplines de la psychologie et des neurosciences.

Le dépassement de l'opposition entre la MMH et MRH dans un pluralisme explicatif intégratif sous le FEP repose sur la relation bidirectionnelle entre aspects fonctionnalistes, qui procède des parties au tout, et structuralistes, qui procède du tout aux parties, dans une architecture cognitive hiérarchiquement organisée. En effet, nous proposons d'envisager l'interdépendance hiérarchique entre les trois thèses ensemble de thèses concurrentes de la façon suivante : (1) le computationnalisme représentationnel et le dynamicisme non-représentationnel sont interdépendants dans une hiérarchie qui comprend d'une part des processus automatiques de type « habituels/pavloviens »[9] aux niveaux inférieurs, ce qui implique à ces niveaux la gestion des affordances, et d'autre part des processus délibérés de type « dirigé vers un but »[10] aux niveaux supérieurs, ce qui implique à ces niveaux la manipulation de représentations; (2) la modularité quasi-décomposable et l'interactivité non-décomposable sont interdépendants dans une hiérarchie qui comprend d'une part des mécanismes « domaine-spécifiques » quasi-décomposables aux niveaux inférieurs, ce qui implique une ségrégation modulaire par des connections denses de courtes portées, et d'autre part des mécanismes « domaine-généraux » aux niveaux supérieurs, ce qui implique une intégration intermodulaire non-décomposable dans des hubs « rich-club » par des connections éparses de longue portée; (3) l'adaptation par évolution mosaïque et l'adaptativité par évolution concertée sont interdépendant dans une hiérarchie qui comprend d'une part des mécanismes « domaine-spécifiques » au neurodéveloppement canalisé aux niveaux inférieurs, ce qui implique une évolution mosaïque qui repose sur des pressions sélectives locales, et d'autre part des mécanismes « domaine-généraux » au neurodéveloppement plastique, ce qui implique une évolution concertée qui repose sur des pressions sélectives globales. Telle est, soutiendrons-nous, la nature de l'architecture cognitive humaine. Dans l'ensemble, il serait ainsi possible d'envisager l'opposition apparente des caractéristiques

[9] Selon la théorie de l'apprentissage par renforcement, les processus habituels (ou automatiques) impliquent la répétition réflexive de séquences d'actions préalablement renforcées. Ce faisant, ils reposent sur l'apprentissage par essais et erreurs (model-free) pour mettre à jour la valeur d'une séquence d'action associée à un stimulus (Daw et al., 2005). Nous pouvons aussi inclure les processus de type « pavloviens » dans les processus automatiques (model-free) puisqu'ils sont associés aux comportements réflexes d'approche et d'évitement. Contrairement aux processus « habituels », ceux-ci reposent sur des priors adaptatifs qui sont prescrits par l'histoire évolutionnaire.

[10] Selon la théorie de l'apprentissage par renforcement, les processus dirigés vers des buts (ou délibérés) impliquent l'évaluation réflective de séquences d'actions alternatives en fonction des conséquences potentielles. Ce faisant, ils reposent sur l'apprentissage par essais et erreurs vicariants (model-based) qui simule des séquences d'actions alternatives suivant un arbre de décision qui associe des transitions d'états et des conséquences potentielles (Daw et al., 2005).

architecturales fonctionnalistes proposées par la MMH et structuralistes proposées par la MRH comme étant complémentaires sous le FEP dans le contexte d'une architecture cognitive hiérarchique.

En récapitulant, ce mémoire contient deux chapitres. Le premier aborde dans la première section la question du pluralisme explicatif intégratif dans la perspective d'une distinction entre explication fonctionnaliste et structuraliste des systèmes biologiques. Dans les autres sections de ce chapitre, nous présenterons le paradigme architectural cognitiviste, détaillant comment les trois hypothèses complémentaires de la MMH (Carruthers, 2006; 2008) – c'est-à-dire le computationnalisme représentationnel, la modularité quasi-décomposable et l'adaptation par évolution mosaïque – mettent l'accent sur la sélection locale de composantes dans l'architecture cognitive. Nous présenterons ensuite le paradigme architectural de la cognition incarnée, détaillant comment les trois hypothèses interdépendantes de la MRH (Anderson, 2014, 2016) – c'est-à-dire le dynamicisme non représentationnel, l'interactivité non-décomposable et l'adaptativité par évolution concertée – mettent l'accent sur le processus d'auto-organisation globale dans l'architecture cognitive. Quant au deuxième chapitre, il argumente en faveur d'un pluralisme explicatif intégratif sous le FEP qui permet, nous l'espérons, de dépasser de l'opposition dichotomique entre les hypothèses architecturales. Pour ce faire, nous ferons appel au FEP pour démontrer comment les hypothèses du cognitivisme classique et de la cognition incarnée concernent en fait des aspects interdépendants de l'architecture cognitive. Après avoir introduit plus en détail le FEP et les théories qui en découlent dans la première section, nous tenterons de démontrer l'interdépendance des thèses qui caractérisent les hypothèses architecturales dans trois sections subséquentes, chacune dédiée à l'une de ces interdépendances. Ainsi la deuxième section s'attarde à l'interdépendance hiérarchique entre computationnalisme représentationnel et dynamicisme non-représentationnel, la troisième section s'attarde à l'interdépendance hiérarchique entre modularité quasi-décomposable et interactivité non-décomposable, et finalement, la quatrième section s'attarde à l'interdépendance hiérarchique entre adaptation par évolution mosaïque et adaptativité par évolution concertée. En guise de conclusion, j'aborderai les implications de la proposition dans d'autres problématiques, notamment la problématique de la conscience.

CHAPITRE 1

PARADIGMES ANTAGONISTES DE L'ARCHITECTURE COGNITIVE : LA MODULARITÉ MASSIVE ET LE REDÉPLOIEMENT MASSIF

Dans ce chapitre, nous abordons les thèses principales qui caractérisent les architectures cognitives que nous avons choisis pour représenter respectivement le cognitivisme et la cognition incarnée, soit l'hypothèse de la modularité massive (MMH) de Carruthers (2006; 2008) et l'hypothèse du redéploiement massif (MRH) d'Anderson (2014; 2016). Dans la section 1.1, nous proposerons d'envisager les architectures cognitives de la MMH et de la MRH dans la perspective d'un pluralisme explicatif intégratif qui distingue entre les conceptions fonctionnalistes et structuralistes de l'explication biologique. Cette distinction permettra de mieux contextualiser le débat entre les architectures cognitives puisque nous soutiendrons qu'elles concernent des aspects qui s'expliquent différemment, mais qui sont ultimement complémentaires. Dans la section 1.2, nous aborderons la MMH dans la perspective d'un fonctionnalisme sélectionniste qui soutient que l'architecture cognitive repose sur des modules fonctionnellement spécialisés qui ont été façonnés par des pressions sélectives locales pour la résolution de problèmes adaptatifs dans l'environnement évolutionnaire. Plus particulièrement, la MMH repose essentiellement sur trois thèses « componentielles » complémentaires que nous aborderons l'une à la suite de l'autre dans ce chapitre: le computationnalisme représentationnel, la modularité quasi-décomposable et l'adaptation par évolution mosaïque. Dans la section 1.3, nous aborderons la MRH dans la perspective d'un structuralisme développementaliste qui soutient que l'architecture cognitive repose sur des biais corticaux fonctionnellement différenciés qui ont été façonnés par des pressions sélectives globales pour permettre le redéploiement des ressources neuronales dans un environnement largement conservé entre les générations. Plus particulièrement, la MRH repose essentiellement sur trois thèses « processuelles » que nous aborderons l'une à la suite de l'autre dans ce chapitre : le dynamicisme non-représentationnel, l'interactivité non-décomposable et l'adaptativité par évolution concertée. Dans l'ensemble, ce chapitre a pour objet d'expliciter les thèses principales des architectures cognitives internalistes et externalistes, ce qui nous permettra de mieux comprendre leurs forces et leurs faiblesses.

1.1 Pluralisme explicatif intégratif : les paradigmes fonctionnaliste et structuraliste pour expliquer l'architecture cognitive

Avant d'aborder les thèses principales qui caractérisent les architectures cognitives, nous proposons de clarifier la forme que prendra la proposition actuelle. Comme nous venons de voir dans l'introduction (et détaillerons dans le deuxième chapitre), nous envisageons la MMH et la MRH comme étant complémentaires dans la perspective d'un pluralisme explicatif intégratif sous le principe d'énergie libre qui distingue entre fonctionnalisme sélectionniste, qui procède des « parties au tout », et structuralisme développementaliste, qui procède du « tout aux parties ». En ce sens, la MMH et la MRH concernent des aspects mutuellement interdépendants dans l'architecture cognitive du cerveau, c'est-à-dire que c'est la relation bidirectionnelle entre les aspects componentiels et processuels qui vont façonner l'architecture cognitive. Pour une compréhension adéquate de l'architecture cognitive, nous soutiendrons que ce pluralisme explicatif intégratif se manifeste dans les quatre niveaux d'analyse de Tinbergen (1963). Nous soutiendrons que, sous une théorie des systèmes évolutionnaires, ces niveaux d'analyse correspondent aux différentes échelles temporelles sur lesquelles la sélection naturelle et l'auto-organisation interagissent (Badcock, 2012; Badcock et al., 2019).

Nous soutiendrons que la sélection naturelle et l'auto-organisation sont mutuellement nécessaires pour une explication adéquate des systèmes biologiques (Kauffman, 1993; Batten et al., 2008; Badcock, 2012; Linde Medina, 2016). Pour ce faire, nous proposons une distinction supplémentaire entre les niveaux sélectionnistes (fonctionnalisme centré sur la fonction biologique), qui soutiennent une conception mécaniste d'une matière biologique passive façonnée de manière extrinsèque par la sélection naturelle, et les niveaux développementalistes (structuralisme centré sur la forme biologique), qui soutiennent une conception structurelle d'une matière biologique active façonnée de manière intrinsèque par l'auto-organisation développementale (Linde Medina, 2010; Barbieri, 2012). Pour le principe d'énergie libre, la sélection naturelle et l'auto-organisation sont deux façons de minimiser l'énergie libre du phénotype de l'organisme, soit par la sélection naturelle de modèles bayésiens (Campbell, 2016) ou soit par l'optimisation de modèles bayésiens (Friston, 2010). En effet, les modèles (ou phénotypes) s'auto-organisent continuellement pour minimiser l'énergie libre et ceux avec l'énergie libre moyenne la plus faible (ou la fitness adaptative la plus élevée) dans une génération sont sélectionnés. Dans l'architecture cognitive, nous soutiendrons que la MMH et la MRH s'attardent à des aspects complémentaires dans la perspective de l'esprit hiérarchiquement mécaniste, c'est-à-dire que la sélection naturelle façonne les régions sensorimotrices hautement ségréguées et canalisées qui fournissent des ancrages

neurodéveloppementaux qui permettent l'auto-organisation progressive des régions associatives hautement intégrées et plastiques au cours du développement (Badcock, 2012; Badcock et al., 2019).

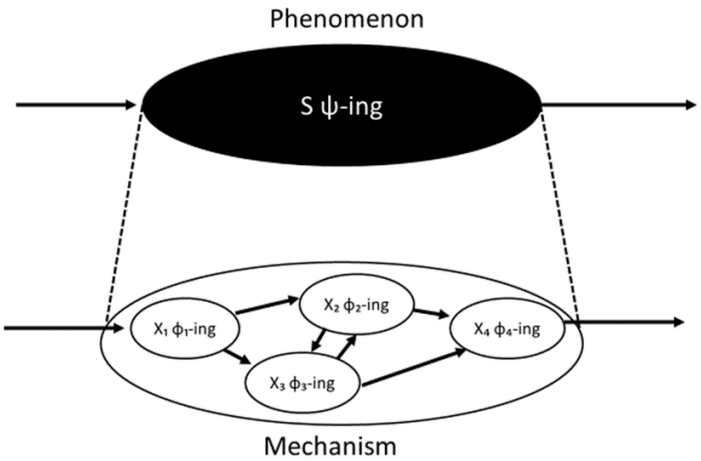
D'un point de vue biologique, nous pouvons comprendre la MMH comme une explication fonctionnaliste de l'architecture cognitive puisqu'elle soutient que la forme dépend de la fonction. Selon ce courant de pensée en biologie, qui persiste malgré les nombreux changements théoriques, la matière biologique est passive (elle reçoit la forme) et n'est pas intrinsèquement organisée. Son organisation requiert l'action d'une force externe (fonction qui façonne les traits) ou d'un facteur organisateur (sélection naturelle) pour expliquer la forme de l'organisme (Linde Medina, 2010). Cette conception mécaniste de la matière biologique découle de l'application de la mécanique newtonienne pour l'étude des systèmes biologiques, ce qu'on associe typiquement aux courants préformationnistes et darwinistes du 18^{e} et 19^{e} siècles, respectivement. Selon cette conception mécaniste, les organismes sont compris dans une perspective componentielle qui procède « des parties au tout », c'est-à-dire qu'ils sont constitués de parties distinctes dont les interactions causales produisent un tout fonctionnel. Dans sa forme contemporaine, le fonctionnalisme est associé à la synthèse moderne qui soutient que la forme est façonnée par la sélection naturelle pour produire des composantes fonctionnelles qui sont adaptées à l'environnement externe (des traits phénotypiques, notamment). Pour expliquer la génération de la forme biologique à partir de la matière biologique passive, les fonctionnalistes contemporains utilisent la métaphore d'un programme développemental sélectionné dans lequel le génome joue un rôle de régulation (Carroll, 2005). Ce programme développemental repose essentiellement sur un mécanisme de réplication moléculaire qui est essentiel à l'hérédité et la sélection naturelle (Barbieri, 2012). En effet, un système d'ARN/ADN auto-répliquant est nécessaire pour permettre à l'organisme de s'adapter à son environnement en reproduisant sélectivement les modifications de son génotype qui s'avèrent avantageuses. Pour le fonctionnalisme, ce sont surtout les contraintes fonctionnelles qui façonnent la forme de l'organisme, c'est-à-dire que les changements évolutionnaires sont guidés par les exigences fonctionnelles de l'environnement externe (Montgomery, Mundy et Barton, 2016). Les contraintes fonctionnelles favorisent ainsi les explications externes qui reposent sur les pressions sélectives qui favorisent l'évolution spécifique de chacune des composantes fonctionnelles, ce qu'on nomme l'hypothèse de l'évolution mosaïque (Barton & Harvey, 2000). Par son emphase sur les contraintes fonctionnelles dans l'évolution

de l'architecture cognitive, la MMH se focalise sur le mécanisme (causation) et la fonction (adaptation), mais néglige les aspects structuraux dans l'explication des autres niveaux d'analyse biologique.

Dans la perspective du fonctionnalisme sélectionniste, les contraintes fonctionnelles vont favoriser l'émergence d'une architecture cognitive componentielle mécaniste. En effet, nous avons vu que pour les fonctionnalistes les changements évolutionnaires sont essentiellement fonctionnels et que l'organisme peut être partitionné de façon à associer chacune des parties à des fonctions. Comme nous le verrons prochainement, la MMH est une architecture cognitive fonctionnaliste paradigmatique avec ses dynamiques à « dominance componentielle » dans lesquelles les propriétés fonctionnelles des parties déterminent les propriétés fonctionnelles du tout (Van Orden, Holden et Turvey, 2003). En d'autres termes, cette architecture cognitive est un système quasi-décomposable puisque les interactions dans les composantes sont plus importantes que les interactions entre les composantes (Simon, 1991). Puisque les systèmes quasi-décomposables comportent des composantes distinctes, celles-ci peuvent être analysées par une stratégie explicative mécaniste (Bechtel & Richardson, 2010). Sommairement, un mécanisme comprend des composantes (ou entités) dont les interactions (ou activité) sont organisées de façon à produire une fonction (ou phénomène explanandum). L'emphase est sur les relations de constitution componentielle, c'est-à-dire que l'explication procède de façon « bottom-up », des « parties au tout », puisqu'on explique un phénomène en se référant aux entités et activités situées sur un niveau inférieur d'organisation. Plus concrètement, l'explication componentielle-mécaniste se déroule par la décomposition des fonctions en opérations et des structures en composantes pour ensuite associer les

deux par la localisation des mécanismes (ce qu'on appelle l'heuristique de décomposition et localisation mécaniste; voir Bechtel & Richardson 2010). Dans une représentation visuelle canonique (figure 1) :

Figure 1 : Représentation d'une explication mécaniste d'un phénomène



Au sommet se trouve le phénomène expliqué : la performance d'un comportement ψ par un système S. En dessous se trouve le mécanisme responsable du comportement de S, soit les composantes du mécanisme (les X) et leurs activités au sein du mécanisme (les ϕ), lesquelles sont organisées selon les flèches liant les X. Les lignes pointillées verticales servent à montrer les parties et les activités qui sont des composantes du mécanisme engagé dans ce comportement. En ce sens, les mécanismes sont décomposables puisque le comportement du système dans son ensemble peut être décomposé en interactions organisées entre les activités des parties. (Tirée de Craver, 2007, p.7)

La stratégie explicative componentielle mécaniste rencontre toutefois un certain nombre de difficultés lorsque vient le temps d'expliquer certaines propriétés systémiques globales, notamment celles qui sont expliquées par les sciences de la complexité. Les sciences de la complexité étudient comment de larges collections de composantes qui interagissent localement peuvent spontanément s'auto-organiser pour exhiber des structures globales. Toutefois, il n'est pas certain qu'on puisse expliquer les propriétés systémiques par une simple agrégation de composantes en interaction linéaire (additive), comme le suppose l'explication causale mécaniste. C'est pourquoi des stratégies explicatives alternatives sont souvent employées dans les sciences de la complexité, notamment la stratégie explicative structuraliste[11]

[11] Cette stratégie explicative implique les nouvelles techniques mathématiques dérivées de la théorie des graphes qui permettent la modélisation des réseaux et de leurs dynamiques. En d'autres termes, la modélisation des réseaux permet la prédiction des comportements globaux des systèmes dynamiques complexes en raison de

pour laquelle ce sont les structures mathématiques (topologie des réseaux, tendance vers un point d'équilibre, etc.) qui jouent un rôle explicatif (Huneman, 2018).

D'un point de vue biologique, nous proposons ainsi de comprendre la MRH dans la perspective d'une explication structuraliste de l'architecture cognitive puisqu'elle soutient que la fonction dépend de la forme. Contre le fonctionnalisme, le courant de pensée structuraliste soutient que la matière biologique est active (génère spontanément la forme) et capable d'auto-organisation, c'est-à-dire qu'il s'agit d'un médium excitable capable d'exhiber de l'ordre intrinsèque par l'interactions entre composantes, sans requérir de facteurs d'organisation externe pour expliquer la forme de l'organisme (Linde Medina, 2010). Cette conception organiciste (ou holiste) de la matière biologique précède historiquement le darwinisme. On l'associe typiquement au courant morphologiste du 19^{e} siècle qui s'est élaboré sous l'influence de Kant, notamment (voir Huneman, 2006). Selon cette conception structuraliste, les organismes sont compris dans une perspective non-componentielle qui procède du tout aux parties, c'est-à-dire qu'ils forment un tout autoorganisé qui contraint les interactions entre parties, lesquelles sont ainsi « causes et effets d'elles-mêmes » (Kant, 1987, p. 249). Dans sa forme contemporaine, le structuralisme est associé à la synthèse étendue développementaliste qui soutient que la forme est façonnée par l'auto-organisation en fonction de principes organisationnels internes pour produire un tout fonctionnel intégré (Kaufmann, 1993). Pour expliquer la génération de la forme biologique à partir de la matière biologique active, les structuralistes contemporains se réfèrent souvent à l'idée d'un système developmental autoorganisé dans lequel le génome n'est qu'un élément interactif parmi d'autres (Oyama et al., 2001). Ce système développemental repose essentiellement sur un processus d'auto-assemblage qui permet l'agrégation spontanée des composantes dans des structures supramoléculaires (Barbieri, 2012). Un exemple paradigmatique d'auto-assemblage est la formation spontanée des membranes de phospholipides dans une solution aqueuse, mais nous pouvons aussi comprendre le processus de développement de structures dans cette perspective (morphogénèse des tissus par auto-assemblage, par exemple). Pour le structuralisme, ce sont surtout les contraintes développementales qui façonnent la forme de l'organisme, c'est-à-dire que les changements évolutionnaires sont guidés par des règles développementales qui vont limiter la portée de la sélection naturelle aux variations permises par les propriétés génératives internes de l'organisme (Montgomery, Mundy et Barton, 2016). Les contraintes développementales favorisent ainsi les explications internes qui reposent sur les pressions sélectives qui favorisent l'évolution par covariation

l'interconnexion non-linéaire (multiplicatives) entre composantes qui génèrent des structures prévisibles, sans reposer sur des composantes mécanistes ou trajectoires causales spécifiques (Moreno & Suarez, 2020).

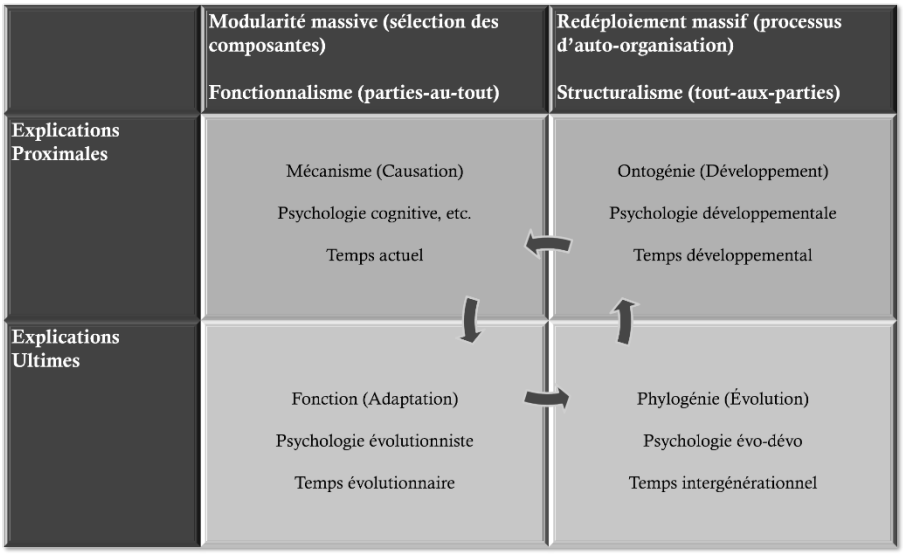
des composantes en fonction d'évènements développementaux, ce qu'on nomme l'hypothèse de l'évolution concertée (Finlay & Darlington, 1995). Par son emphase sur les contraintes développementales (pouvant être comprises positivement comme des contraintes habilitantes[12]; voir Raja & Anderson, 2021) dans l'évolution de l'architecture cognitive, la MRH se focalise sur la phylogénie (évolution) et l'ontogénie (développement), mais néglige les aspects fonctionnels des autres niveaux d'analyse biologique.

Dans la perspective du structuralisme développementaliste, les contraintes développementales vont favoriser l'émergence d'une architecture non-décomposable. En effet, nous avons vu que pour les structuralistes les changements évolutionnaires sont essentiellement des changements de forme et que la morphologie résulte de règles développementales uniformes. Comme nous le verrons prochainement, la MRH est une architecture cognitive structuraliste paradigmatique avec ses dynamiques à « dominance interactive » dans lesquelles les propriétés fonctionnelles des parties sont déterminées par les propriétés fonctionnelles du tout (Van Orden, Holden et Turvey, 2003). Cette architecture cognitive est un système non décomposable puisque les interactions entre les composantes sont plus importantes que les interactions dans les composantes (Van Orden, Holden et Turvey, 2003). Puisque les systèmes non décomposables sont des systèmes dont les composantes sont hautement intégrées, elles résistent à la stratégie explicative mécaniste traditionnelle (Rathkopf, 2018). En guise d'alternative, une stratégie explicative qui repose sur des contraintes habilitantes permet de mieux capturer des relations systémiques qui ne sont pas comprises dans les explications componentielles mécanistes (Anderson, 2015). Sommairement, les contraintes habilitantes limitent le degré de liberté d'un système pour lui permettre de réaliser une activité qu'il lui serait impossible à réaliser autrement. Plus formellement (Anderson, 2015): « Enabling constraint = A physical relationship between a functional system S and entities {X} (and/or mechanism M), at the same or different level of description, such that {X} (and/or M) changes the relative probabilities of various possible functional outcomes of activity in S » (p. 12). Les contraintes habilitantes permettraient d'expliquer des phénomènes particuliers tels que l'activité des cellules amacrines dans la rétine des mammifères dont les dendrites sont caractérisées par une fonction de détection de mouvement qui est spécifique à une direction, laquelle serait difficile à comprendre sans considérer la cellule comme une partie d'un système plus important qui contraint l'activité de la cellule de façon à la rendre

[12] Selon Raja et Anderson (2021, p. 217), « a positive understanding of developmental constraints places them among the enabling constraints of evolution. Developmental constraints change the probability of the outcomes of evolutionary processes by actively providing them with a directionality—i.e., with a way in which natural selection itself may affect or not different organisms. In this sense, developmental constraints enable evolutionary processes to be the way they are ».

fonctionnelle (Anderson, 2015). L'emphase est sur des relations qui vont au-delà de la constitution componentielle pour inclure des contraintes habilitantes, c'est-à-dire que l'explication procède aussi de façon « top-down », du « tout aux parties », puisqu'on explique un phénomène en se référant aux contraintes situées sur un niveau supérieur d'organisation. Plus concrètement, l'explication par contrainte habilitante procède en identifiant comment des interactions structurées dans un système permettent aux mécanismes de réaliser certaines de leurs activités.

Figure 2 : Pluralisme explicatif intégratif et niveaux d'analyse de Tinbergen



L'architecture cognitive du cerveau est le produit des interactions complémentaires entre la sélection (générale et naturelle) et de l'auto-organisation sur quatre niveaux d'analyse de Tinbergen. Les explications fonctionnalistes (des parties au tout) focalisent sur la sélection naturelle des composantes locales. Les explications structuralistes (du tout aux parties) focalisent sur l'auto-organisation des processus globaux. Les paradigmes psychologiques se concentrent différemment sur les niveaux d'analyse biologique : 1) hypothèses fonctionnelles pour les caractéristiques adaptatives typiques de l'espèce (c'est-à-dire la psychologie évolutionniste et la synthèse moderne néo-darwinniennne); 2) explications intergénérationnelles des similitudes et différences entre les groupes (c'est-à-dire la psychologie évo-dévo et la synthèse évolutionnaire étendue); 3) explications pour le développement individuel (c'est-à-dire psychologie développementale); et 4) explications

> mécanistes pour les phénomènes biocomportementaux en temps réel (c'est-à-dire les sous-disciplines de la psychologie).

1.2 L'hypothèse de la modularité massive : le cerveau comme système de manipulation de représentations

Dans cette section, nous introduisons les thèses complémentaires qui caractérisent l'hypothèse de la modularité massive (MMH). Selon la MMH fonctionnaliste et sélectionniste, les capacités cognitives humaines résultent de modules computationnels domaine-spécifiques qui sont principalement façonnés par le processus évolutionnaire de sélection naturelle (Sperber, 1994; Cosmides & Tooby, 1994; Pinker, 2003; Barrett & Kurzban, 2006; Carruthers, 2006). Plus particulièrement, trois hypothèses qui caractérisent la MMH : (1) Le cerveau est un système computationnel de manipulation de représentations; (2) Les processus cognitifs résultent de l'opération de, et d'interactions entre, modules computationnels possédant une réalisation neuronale spécifique (dans une association structure-fonction « one-to-one », mais avec certaines réserve[13]); (3) L'architecture fonctionnellement spécialisée a été façonnée par des pressions sélectives locales dans un processus évolutionnaire de sélection naturelle. Puisque la MMH est une proposition radicale qui s'applique à la quasi-totalité de l'architecture cognitive, elle nous permettra de mettre en relief les thèses complémentaires qui la caractérise, soit le computationnalisme représentationnel, la modularité quasi-décomposable et l'adaptation par évolution mosaïque. Dans les sous-sections suivantes, nous aborderons les versions plus récentes de la MMH, notamment celle proposée par Carruthers (2006; 2008).

1.2.1 L'architecture computationnelle

Comme nous l'avons précédemment abordé, la MMH souscrit au paradigme cognitiviste classique[14] qui soutient que la cognition se limite essentiellement à l'activité du système nerveux. Les inputs sensoriels, qui sont associés à la perception, et les outputs moteurs, qui sont associés à l'action, sont seulement compris comme des points d'entrée et de sortie. Sans un accès direct au monde, la cognition est alors

[13] L'hypothèse de la modularité massive soutient une modularité fonctionnelle qui n'implique par une localisation des mécanismes computationnels dans des régions neuronales restreintes. En effet, les mécanismes computationnels peuvent être distribués sur plusieurs régions neuronales (Carruthers, 2006, p. 4).

[14] Nous distinguons entre le cognitivisme (ou classicisme), pour lequel les représentations mentales sont des structures symboliques localisées qui sont manipulées fonction de règles (Fodor, 1975; Fodor et Pylyshyn; 1988, etc.), et le connexionnisme, pour lequel les représentations mentales sont des patrons d'activation distribués d'un réseau de noeuds (Rumelhart & McClelland, 1986; Smolenski, 1988, etc.). En général, les noeuds et patrons d'activation des réseaux connexionnistes ne possèdent pas de contenu sémantique, mais on peut envisager des variantes localistes dans lesquelles les noeuds possèdent des propriétés sémantiques (voir Ballard, 1986).

interprétée comme un processus « orienté vers la perception » qui repose sur la construction puis la manipulation de représentations découplées de l'environnement. En d'autres termes, on comprend la cognition comme un processus reconstructif où la perception a pour fonction de reconstruire, dans des modèles internes, la structure objective du monde. Pour mieux définir cette notion, les cognitivistes associent typiquement la manipulation de représentations au computationnalisme, c'est-à-dire qu'on comprend le fonctionnement de la cognition comme étant analogue à celui d'une machine de Turing[15] (physiquement réalisée) qui manipule des symboles en fonction de règles (voir Eliasmith, 1997). C'est pourquoi les cognitivistes qualifient souvent la cognition comme un processus computationnel de traitement de l'information qui s'apparente (à un niveau abstrait et formel de description) à ce qui se passe dans un ordinateur digital. Il existe toutefois plusieurs façons de comprendre ce traitement de l'information (notamment, en termes d'information de Shannon, qui peut être digitale ou analogue; voir Piccinini & Scarantino, 2011), mais les cognitivistes de réfèrent la plupart du temps à un processus de traitement de l'information sémantique qui implique une manipulation de représentations qui sont « à propos » de quelque chose (Fodor, 1975).

Ce computationnalisme sémantique peut être compris comme la conjonction entre la théorie représentationnelle de l'esprit et la théorie computationnelle de l'esprit. Cette conjonction entre une approche computationnelle de la cognition et une « intentionnalité représentationnelle » ou cognitive fut introduite par Fodor dans le *Langage de la pensée* (1975). Selon la théorie du langage de la pensée (LOT), la pensée (comprise en termes d'attitudes propositionnelles) repose sur un système de représentation qui est une sorte de métalangage qui associe une sémantique compositionnelle avec une syntaxe combinatoire (pour plus de détails sur cette conception de la structure des représentations mentales, voir Fodor & Pylyshyn, 1988). On comprend les attitudes propositionnelles (telles que les croyances, désirs et intentions) comme des états mentaux intentionnels, c'est-à-dire qu'ils sont à propos d'objets du monde. Dès lors, les représentations du LOT sont comprises comme des reconstructions objectives du monde qui expriment des conditions de vérité (ou de satisfaction), c'est-à-dire qu'elles doivent satisfaire des conditions normatives pour être vraies, réalisées ou exécutées (Fodor, 1990). En d'autres termes, les

[15] Une machine de Turing est un modèle mathématique d'une machine informatique hypothétique qui peut utiliser un ensemble prédéfini de règles pour déterminer un output à partir d'un ensemble de variables d'input.

représentations mentales possèdent des propriétés sémantiques qui permettent de représenter correctement ou incorrectement leurs objets dans le monde[16].

Nous pouvons comprendre la MMH comme une perspective radicale puisque la quasi-totalité de l'architecture cognitive serait ainsi computationnelle-représentationnelle. C'est pourquoi nous consacrerons le reste de cette sous-section à présenter le computationnalisme dans cette perspective particulière. En effet, la MMH de Carruthers s'inscrit dans cette perspective orientée vers la perception (ou reconstructive) pour laquelle la plupart des processus cognitifs reposent sur la manipulation de représentations découplées de l'environnement. Pour être exact, ce sont les représentations locales qui sont essentielles pour décrire les capacités cognitives humaines – les représentations globales qu'on associe aux patrons d'activité des réseaux connexionnistes distribués ne sont pas exclus d'emblées, mais selon Carruthers (2006, p.46), ces-dernières ne peuvent jouer qu'un rôle périphérique dans certains processus de bas niveau. C'est parce que dans cette perspective computationnelle-représentationnelle, l'essentiel de l'architecture cognitive humaine s'articule surtout autour d'une variante peu exigeante du LOT, en ce sens que la seule exigence cruciale est une structure componentielle qui implique des représentations locales (Carruthers, 2006, p.46)[17]. Cette variante du computationnalisme repose sur une théorie componentielle de l'esprit, c'est-à-dire que le cerveau opère des transformations algorithmiques sur des représentations locales, compositionnellement structurées et causalement efficaces. Dans une sémantique compositionnelle, la signification des représentations complexes est construite en combinant

[16] Comme nous le verrons, les propriétés sémantiques des représentations (notamment, les conditions de vérité, le contenu et la référence) sont problématiques puisqu'elles sont essentiellement des propriétés mentales. Il faudrait alors proposer une théorie naturalisée du contenu qui s'appuie sur des propriétés physiques, mais plusieurs auteurs jugent qu'il s'agit d'une tâche virtuellement impossible (voir le problème difficile du contenu; Hutto & Myin, 2012). Selon eux, il faudrait alors abandonner le projet de naturalisation et opter pour une forme plus basique d'intentionnalité qui ne requiert pas de contenu sémantique (Hutto & Satne, 2015; Kiverstein & Rietveld, 2015). Dans le second chapitre, nous soutiendrons que ce scepticisme est justifié puisqu'une intentionnalité avec contenu serait inadéquate pour décrire certains processus, notamment les processus que nous appellerons « model-free » ou « habituels/automatiques/pavloviens », lesquels reposent sur la gestion d'interactions dynamiques concrètes, couplées avec l'environnement.

[17] Carruthers s'appuie sur les arguments de Marcus (2001) pour affirmer qu'une transformation algorithmique de représentations compositionnellement structurées est nécessaire pour rendre compte des capacités cognitives humaines. Notamment, Marcus soutient que les représentations globales qui reposent sur des patrons d'activité dans des réseaux connexionnistes distribués seraient incapable de distinguer entre individus et attributs, ce qui serait essentiel pour des capacités comme la permanence de l'objet. Toutefois, on admet que les réseaux connexionnistes distribués peuvent tout de même jouer un rôle dans certains processus cognitifs de bas niveaux qui reposent sur des formes simples d'inférences statistiques (Carruthers, 2006, p.46).

des composantes plus simples en suivant une structure algorithmique, ce qui fait que les représentations complexes sont décomposables en composantes plus simples. Dans cette variante peu exigeante, le LOT peut incorporer des modèles mentaux ou des cartes mentales, pour autant qu'elles satisfont les exigences d'une structure componentielle/compositionnelle (Carruthers, 2006, p.46). Les modèles mentaux sont des représentations (ou simulations) quasi-perceptuelles des relations entre caractéristiques saillantes d'une situation en mémoire de travail, lesquels sont essentiels pour permettre la prise de décision, la planification et le raisonnement contrefactuel (Johnson-Laird, 1983). De même, la navigation allocentrique (centrée sur les objets) dépend de cartes mentales qui reposent sur des relations compositionnellement structurées entre points de repères (pour autant qu'il s'agisse d'une carte avec des « labels »). Dans l'ensemble, un LOT serait essentiel pour générer le contenu propositionnel, causalement efficace, de la pensée, laquelle permet de contrôler et de motiver l'action au cours des cycles de répétition mentale « créative » (notamment, la répétition mentale de phrases de langage naturel dans un discours intérieur chez l'humain).

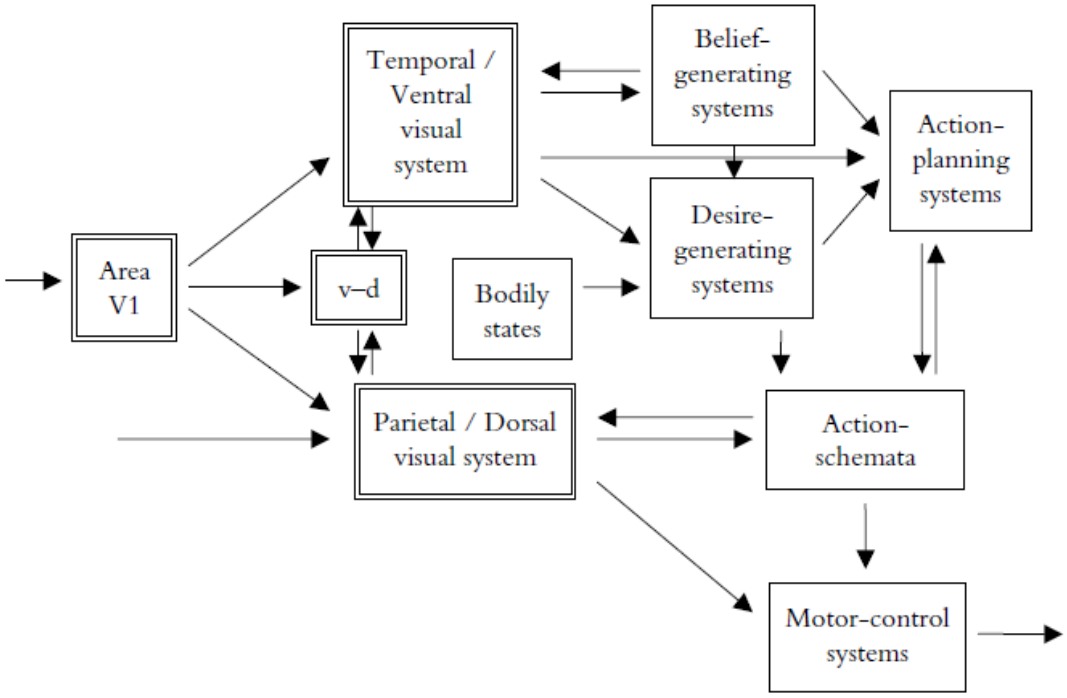
Pour la MMH de Carruthers, la répétition mentale repose toujours sur des boucles sensorimotrices de rétroaction impliquant une médiation représentationnelle, que cette dernière soit faite de représentations perceptuelles non-conceptuelles ou quasi-perceptuelles conceptuelles/propositionnelles. En d'autres termes, les boucles sensorimotrices de rétroaction dépendent d'une séparation entre les représentations (quasi)perceptuelles et les schémas d'action, dans la perspective classique d'un modèle du comparateur qui repose sur des commandes motrices et des copies d'efférences[18] (voir Miall et Wolpert, 1996). Plus précisément, Carruthers distingue entre deux formes de répétition mentale des schémas d'action ayant des fonctions différentes, l'une consciente et l'autre inconsciente (figure 3) (Carruthers, 2006, p.140). Pour la répétition mentale consciente, une diffusion dans l'espace de travail global (dans le sens de Baars, 2007)[19] conféré par la mémoire de travail et la perception serait nécessaire. L'espace de travail global est essentiel selon Carruthers pour permettre des cycles de répétition mentale « créative » qui cooptent les modules centraux pour attribuer des valeurs de contrôle (ou croyances) et de motivation (ou désirs) au contenu propositionnel exprimé dans les représentations d'un LOT. D'un point de vue neurobiologique, la

[18] Dans un modèle du comparateur, lorsqu'un schéma d'action est activé, il génère à la fois des commandes motrices qui permettent de contrôler l'action et des copies d'efférence internes qui génèrent une représentation des perceptions qui sont prédites comme conséquences de l'exécution du schéma d'action (Wolpert et al. 1995).

[19] Selon la théorie de l'espace de travail global (une forme de mémoire de travail et de perception), la fonction première de la conscience serait la diffusion globale de l'information dans des boucles thalamocorticales, ce qui permet de recruter diverses ressources internes pour un traitement plus approfondi (Baars, 2007).

répétition mentale consciente serait associée aux boucles de rétroaction de la voie ventrale/temporale qui génère le contenu quasi-perceptuel dans une perspective allocentrique, centrée sur les objets (Carruthers, 2006, p.84). Il convient de noter que la répétition mentale consciente qui implique la diffusion globale des schémas d'action planifiés est réalisée par un système central virtuel de plus haut niveau, lequel intègre les différents modules de plus bas niveau dans l'espace de travail conféré par la mémoire de travail et la perception (Carruthers, 2006, p. 141). En ce qui concerne maintenant la répétition mentale inconsciente, il n'y a pas dans ce cas de diffusion globale puisque les boucles de rétroaction sont réalisées localement, dans le système dorsal/pariétal. Ici, les copies efférentes des schémas moteurs activés sont rétroprojetées pour rencontrer et pour être comparés avec l'information en provenance des inputs sensoriels, ce qui permet un ajustement rapide et inconscient de l'action experte. D'un point de vue neurobiologique, la répétition mentale inconsciente est associée aux boucles de rétroaction de la voie dorsale/pariétale qui génère un contenu perceptuel (non-conceptuel et continu) qui permet le guidage de l'action dans une perspective égocentrique (centrée sur le sujet) (Carruthers, 2006, p.85). Toutefois, il semble que cette conception intellectualiste de l'action experte, qui nécessite de se représenter à soi-même le schéma d'action par une copie d'efférence, soit inutilement exigeante sur le plan informationnel. Comme nous le verrons dans le second chapitre, nous soutiendrons qu'il est plus parcimonieux de concevoir les boucles de rétroactions de la voie dorsale/pariétale comme un processus d'ajustement dynamique qui repose sur la gestion d'interactions sensorimotrices qui sont directement couplées avec l'environnement.

Figure 3 : Les deux voies de la répétition mentale chez Carruthers



Le système visuel est divisé en système dorsal/pariétal et système ventral/temporal. La voie dorsale (située en bas dans le schéma) permet la répétition mentale inconsciente par une boucle de rétroaction directe, qui ne transige pas par les systèmes de croyances, de désirs et de planification de l'action. La voie ventrale (située en haut dans le schéma) permet la répétition mentale consciente par une boucle de rétroaction indirecte, qui transige par les systèmes de croyances, de désirs et de planification de l'action (tirée de Carruthers, 2006, p.92).

1.2.2 L'architecture modulaire

Cette conception computationnelle-représentationnelle de l'architecture cognitive implique un certain nombre de conséquences théoriques. Puisqu'on comprend la cognition comme un processus complexe de manipulation de représentations riches en contenu informationnel, lequel s'appuie uniquement sur des ressources internes du système nerveux, on suppose qu'une subdivision du processus de traitement de l'information dans des modules computationnels est nécessaire pour réduire sa complexité. Dans cette perspective, l'architecture cognitive est comprise comme un système computationnel quasi-décomposable. Dans un système quasi-décomposable, les dynamiques sont à dominance componentielle, c'est-à-dire que les interactions dans les composantes sont plus importantes que les interactions entre les composantes (Simon, 1991). Dans *The Architecture of Complexity*, Simon impose deux conditions pour la quasi-décomposabilité (1991, p. 210) :

> (1) The short-run behavior of each of the component subsystems is approximately independent of the short-run behavior of the other components.
> (2) In the long run, the behavior of any one of the subsystems depends on the behavior of the other subsystems only in an aggregate way (p.210).

Dans un système hiérarchiquement quasi-décomposable, les sous-systèmes atteignent leurs points d'équilibres d'une façon relativement indépendante à court terme, mais les sous-systèmes convergent vers un point d'équilibre commun à long terme. Plus particulièrement, nous pouvons distinguer trois propriétés importantes pour la quasi-décomposabilité (voir Bechtel et Richardson, 2010). Premièrement, l'indépendance entre différents sous-systèmes varie en fonction de l'échelle temporelle considérée. En d'autres termes, les sous-systèmes ne sont pas totalement indépendants, mais relativement indépendants lorsque considérés sur une échelle temporelle suffisamment courte. Deuxièmement, la quasi-décomposabilité permet d'aborder les sous-systèmes en fonction de leurs points d'équilibre locaux. Bien que les sous-systèmes tendent toujours vers un équilibre commun à long terme, ce qui se passe dans les sous-systèmes demeure plus important que ce qui se passe entre eux puisqu'ils atteignent leurs points d'équilibre indépendamment les uns des autres. Troisièmement, la quasi-décomposabilité n'implique pas une décomposabilité complète, la dépendance agrégative entre les différents sous-systèmes est aussi importante que l'indépendance.

Dans une perspective computationnelle, la notion de quasi-décomposabilité est typiquement associée à la notion de modularité. C'est d'ailleurs en s'inspirant fortement des travaux de Simon sur la quasi-décomposabilité que Marr élabore son principe de conception modulaire (1976) :

> Any large computation should be split up and implemented as a collection of small sub-parts that are as nearly independent of one another as the overall task allows. If a process is not designed in this way, a small change in one place will have consequences in many other places. This means that the process as a whole becomes extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous compensating changes elsewhere. (p. 485)

Dans cette conception, les interactions entre composantes sont faibles, ce qui permet de comprendre l'organisation comme étant, dans une première approximation, modulaire. Raffinant le modèle de la vision de Marr, Fodor propose une conception plus explicite et exigeante de la modularité dans *The Modularity of Mind* (1983). Selon Fodor, une architecture cognitive est (périphériquement) modulaire lorsqu'elle

possède, *dans une mesure suffisante*, la plupart des neuf caractéristiques typiquement co-occurentes qu'il identifie[20]. Parmi les caractéristiques présentées, la « spécificité du domaine » et « l'encapsulation informationnelle » sont jugées plus importantes que les autres, cette dernière étant considérée comme étant la caractéristique essentielle de la modularité (Fodor, 2001). Brièvement, la spécificité de domaine indique que les modules sont spécialisés pour opérer sur des classes d'informations spécifiques tandis que l'encapsulation informationnelle indique que l'accès à l'information se trouvant à l'extérieur du module est limitée (Fodor 1983). Paradoxalement, étant donné le titre de son livre, les processus centraux associés à la fixation des croyances ne peuvent pas selon Fodor être réalisés dans des modules encapsulés puisqu'il s'agit de processus globaux (qui sont, selon lui, quinéens et isotropes[21]). Seuls les processus périphériques associés aux récepteurs et aux effecteurs sont des modules computationnels tels qu'il les définit (Fodor, 2001). Toutefois, la conception de modularité mentale proposée par Fodor exclut d'emblée une explication computationnelle des processus centraux, ce qui a motivé les partisans de la MMH à la reconceptualiser dans une perspective fonctionnelle en termes de « mécanismes fonctionnellement spécialisés », pour pouvoir l'étendre aux processus centraux (Cosmides et Tooby, 1994).

On peut envisager la MMH comme une perspective radicale qui pousse cette conception modulaire quasi-décomposable de l'architecture cognitive jusqu'à ses limites conceptuelles. C'est pourquoi nous consacrerons le reste de cette sous-section à présenter la modularité dans cette perspective particulière. Dans l'ensemble, la MMH soutient que la quasi-totalité des processus cognitifs émergent de l'opération de, et de l'interaction entre, modules fonctionnellement spécialisés dont les opérations internes sont largement inaccessibles aux autres sous-systèmes. Les modules de Carruthers sont profondément ancrés dans une conception biologique qui repose sur la notion de quasi-décomposabilité de Simon (voir Milkowski, 2009 pour cette analyse qui soutient que les modules cognitifs sont un sous-type de modules biologiques). En ce sens, il s'agit d'une conception particulièrement faible de modularité qui s'apparente à celle du sens commun, comprise en termes de composantes fonctionnelles distinctes (Carruthers, 2006; voir aussi Barrett & Kurzban, 2006). En effet, pour Carruthers (2006, p. xii): « The result is a notion of

[20] Les neuf caractéristiques typiquement coocurentes proposées par Fodor (1983) sont: (1) la spécificité du domaine; (2) déclanchement obligatoire; (3) l'impénétrabilité cognitive; (4) la rapidité de traitement; (5) l'encapsulation informationelle; (6) l'architecture neuronale fixe; (7) le dysfonctionnement spécifique; (8) la superficialité des outputs; et (9) le programme ontogénique caractéristique.

[21] Selon Fodor (1983), la cognition centrale est caractérisée par deux propriétés qui s'opposent à la modularité: Elle est isotropique, c'est-à-dire qu'elle peut construire des hypothèses sur la base de toutes les connaissances disponibles, et elle est quinnienne, c'est-à-dire que le degré de confirmation associé à une hypothèses dépend de l'entièreté du système de croyances.

modularity that is some distance from Fodor's (in particular, modules needn't be informationally encapsulated). It is much closer to the use of the term 'module' in biology, and it is even closer to the notion used by researchers in artificial intelligence ». Carruthers reprend d'ailleurs l'argument de Simon pour défendre la quasi-décomposabilité hiérarchiques des systèmes fonctionnels complexes dans « the argument from design » (Carruthers, 2006) :

> So Simon's argument is really an argument from design, then, whether the designer is natural selection (in the case of biological systems) or human engineers (in the case of computer programs). It predicts that, in general, each element added incrementally to the design should be realized in a functionally distinct sub-system, whose properties can be varied independently of the others (to a significant degree, modulated by the extent to which component parts are shared between them). It should be possible for these elements to be added to the design without necessitating changes within the other systems, and their functionality might be lost altogether without destroying the functioning of the whole arrangement. And since there are many ancient and evolutionarily significant capacities of the human mind (as well as many capacities constructed by learning of various sorts), we should expect the human mind to be massively modular in its organization (using the weak sense of 'module') (p. 25-26).

Dans cette perspective, le sens faible de module correspond aux composantes « kludgy »[22] (voir Marcus, 2008) d'un système biologique complexe hiérarchiquement quasi-décomposable qui peuvent se superposer partiellement, mais qui demeurent fonctionnellement distinctes. Par conséquent, on peut s'attendre à ce que les modules hiérarchiquement quasi-décomposables soient eux-mêmes composés de sous-modules et à ce que plusieurs modules partagent (ou réutilisent) des sous-modules (Carruthers, 2006, p.60).

Selon Carruthers (2006), les modules fonctionnellement spécialisés sont pour la plupart des mécanismes computationnels « domaines-spécifiques » avec une réalisation neuronale spécifique (bien que parfois distribuée sur plusieurs régions dans une association structure-fonction « one-to-one »). Puisque les modules computationnels sont des sous-systèmes quasi-décomposables qui sont relativement indépendants du reste du système, ils ne peuvent pas être compris comme des modules complètement encapsulés au sens traditionnel ou « narrow-scope » (voir Fodor, 2000). En fait, Carruthers retourne

[22] Un « kluge » est une solution improvisée à un problème, maladroitement assemblée à partir de n'importe quel matériau immédiatement disponible. Marcus (2008) soutient que le cerveau humain utilise beaucoup de kluges, et que la psychologie évolutionniste favorise souvent les gènes qui donnent des avantages immédiats par rapport aux gènes qui fournissent des avantages à long terme.

l'argument de Fodor en faveur de l'encapsulation informationnelle des processus périphériques contre lui-même dans « the argument from computational tractability » (Carruthers, 2006) :

(1) The mind is computationally realized.
(2) All computational mental processes must be suitably tractable.
(3) Only processes that are informationally encapsulated are suitably tractable.
(4) So the mind must consist entirely of encapsulated computational systems.

En effet, l'argument de la tractabilité computationnelle permet à Fodor de soutenir que la cognition centrale ne peut pas reposer sur des modules computationnels, mais cet argument repose sur une encapsulation « narrow-scope » qui implique l'inaccessibilité complète de l'information externe. Pour Carruthers (2006 p. 53), l'argument implique seulement que processus computationnels soient suffisamment frugaux dans leur utilisation de l'information et dans la complexité de leurs algorithmes. On peut satisfaire cette contrainte de frugalité par exemple au moyen d'heuristiques de recherche qui permettent d'accéder à l'information stockée en mémoire, combinée à des règles d'arrêt si la recherche échoue dans des délais prescrits. Dans cette perspective, il est possible d'envisager une modularité computationnelle moins exigeante, laquelle s'articule autour d'une encapsulation faible ou « wide-scope » qui soutient que les modules ne sont pas affectés par la plupart de l'information disponible au cours de leurs opérations, ce qui serait suffisant pour satisfaire la contrainte de frugalité (Carruthers 2006 p.58). On peut toutefois s'interroger sur la robustesse, ou même la pertinence, d'une conception aussi faible de la modularité, mais cette inquiétude peut être dissipée lorsqu'on prend en considération l'ensemble des arguments présentés jusqu'ici en sa faveur. En effet, la modularité de Carruthers est une conception architecturale fondée sur la notion solide de quasi-décomposabilité, c'est-à-dire qu'elle implique davantage qu'une analyse fonctionnelle en des termes purement théoriques et spéculatifs. En ce sens, la fonctionnalité est dérivée de la structure et de la localisation des sous-systèmes modulaires dans l'organisation du système et pas l'inverse. Ce qui rend la notion de modularité de Carruthers robuste et intéressante, c'est qu'il s'agit de sous-systèmes computationnels quasi-décomposables qui peuvent être analysés par heuristique de décomposition et localisation mécaniste[23]. Toutefois, il semble qu'on doit s'attendre à ce que plusieurs mécanismes soient capables de réaliser plusieurs fonctions, comme c'est

[23] Rappelons que cette heuristique, qui décompose les fonctions en opérations qui les composent et les structures en composantes qui les composent, ce qui permet d'associer les deux lors de la localisation (Craver 2007; Bechtel et Richardson 2010).

souvent le cas dans les systèmes biologiques. En excluant d'emblée la présence de mécanismes multifonctionnels (commun dans le monde biologique) ou domaine-généraux (puisqu'on suppose qu'ils ne peuvent pas avoir évolué par sélection naturelle) la MMH se prive inutilement d'outils qui pourrait faciliter l'explication de l'intégration fonctionnelle « sensible au contexte » qui supporte les processus conscients. Dans le second chapitre, nous soutiendrons que des mécanismes neurocognitifs « domaine-généraux » sont nécessaires pour permettre l'intégration intermodulaire par des connexions éparses de longue portée.

1.2.3 L'architecture adaptée

Il nous reste maintenant une dernière thèse importante à aborder dans notre survol des thèses principales de cette version du paradigme cognitiviste. Dans la mesure où la cognition est un processus reconstructif qui repose sur des représentations riches en contenu informationnel, les cognitivistes vont se retrouver avec un problème de sous-détermination de l'information. Selon l'argument de la pauvreté du stimulus (Chomsky, 2014), l'information en provenance de l'environnement développemental serait trop appauvrie pour permettre l'apprentissage de capacités cognitives typiquement humaines. C'est pourquoi on va postuler l'existence de modules qui, pour remplir leur fonction, doivent dépendre d'informations génétiquement héritées pour compenser cette sous-détermination de l'expérience. Sur la base de ce type d'argument, les cognitivistes vont postuler toutes sortes de modules d'apprentissage avec spécification innée pour expliquer l'acquisition du langage (Chomsky, 1975; Fodor, 1983), la théorie de l'esprit (Fodor, 1983; Baron-Cohen, 1997), etc. Selon cette perspective, l'acquisition du langage ou de la théorie de l'esprit implique des modules qui dépendent d'une spécification innée. En revanche, il existe plusieurs façons (souvent contradictoires[24]) de comprendre la spécification innée des modules. Selon Fodor (1983), la spécification innée des modules est comprise comme un programme ontogénique caractéristique qui répond aux déclencheurs environnementaux. Toutefois, la spécification innée des modules acquerra une tout autre signification lorsqu'on l'envisage dans une perspective néo-darwinienne. Dans cette perspective, les modules innés sont des adaptations qui résultent du processus de sélection naturelle dans l'environnement évolutionnaire.

[24] Il existe une panoplie de définition de l'innéisme (absence d'apprentissage, origine génétique des traits, robustesse développementale, héritabilité significative, etc.) qui entretiennent des relations complexes qui sont parfois antagonistes (Mameli & Bateson, 2011).

Pour les partisans de la psychologie évolutionniste néo-darwinienne, la sélection naturelle est la seule force capable d'expliquer l'évolution de systèmes fonctionnels complexes comme le cerveau (Tooby & Cosmides, 2015). Dans cette perspective, l'architecture cognitive humaine résulterait d'un processus de sélection naturelle qui aurait progressivement façonné des modules computationnels spécialisés dans la résolution de problèmes adaptatifs dans l'environnement évolutionnaire. Cette conception de l'évolution des systèmes fonctionnels complexes s'articule autour de la contrainte de modifiabilité séparée des computations biologiques (voir Sternberg, 2011), laquelle est implicitement présente dans des nombreux arguments heuristiques présentés jusqu'alors qui militent en faveur d'une architecture cognitive hiérarchiquement modulaire (Simon, 1991; Marr, 1982; Carruthers, 2006). Par exemple, Simon soutenait notamment que la quasi-décomposabilité est une propriété importante pour l'évoluabilité des systèmes complexes (Simon, 1991). Dans cette perspective, l'évoluabilité des systèmes fonctionnels complexes implique des pressions sélectives locales sur les composantes fonctionnelles, ce qui permet une canalisation innée qui façonne des modules adaptés à l'environnement ancestral. En biologie évolutionnaire, cette conception de l'évoluabilité, qui s'articule autour de la modifiabilité séparée, correspond à l'hypothèse de l'évolution mosaïque. Cette hypothèse implique des pressions sélectives locales qui agissent sur des régions spécifiques, sans entraîner de réponses adaptatives d'autres régions du cerveau (Barton et Harvey, 2000). Dans cette perspective, si le volume (ou toutes autres propriétés) d'un organe évolue indépendamment du volume des autres organes, son évolution peut être considérée comme étant mosaïque. L'hypothèse de l'évolution mosaïque du cerveau bénéficierait d'un certain degré de confirmation empirique, notamment en ce qui concerne la taille relative de certaines régions par rapport à certains aspects du comportement et de l'environnement (Barton & Harvey, 2000; Hager et al., 2012). Dans la littérature, on associe souvent l'hypothèse de l'évolution mosaïque aux contraintes fonctionnelles locales, c'est-à-dire qu'on suggère que les composantes tendent à évoluer séparément puisque les pressions sélectives agissent sur des composantes fonctionnellement connectées, possiblement pour renforcer des comportements spécialisés (Barton and Harvey, 2000; Montgomery et al. 2016).

On peut comprendre la MMH comme une perspective radicale puisqu'elle soutient que la quasi-totalité de l'architecture cognitive repose sur une spécialisation fonctionnelle qui est adaptée à l'environnement évolutionnaire, mais pas à des environnements différents (hormis quelques « spandrels », c'est-à-dire des

sous-produits d'adaptations). La MMH implique une canalisation innée[25] significative qui repose sur un programme développemental qui génère des modules « domaines-spécifiques » par des interactions gènes-environnement complexes (Carruthers, 2006, p.21). Ce programme développemental permet d'expliquer comment un développement fonctionnel robuste peut émerger d'interrupteurs génétiques qui activent ou désactivent le développement de modules spécialisés (Carruthers, 2006, p.14). Le programme développemental construit d'ailleurs plusieurs modules innés qui sont aussi des modules d'apprentissage spécialisés[26], lesquels ont été façonnés par l'évolution pour générer des modules acquis au cours du développement (des modules d'acquisition d'habiletés, par exemple) (Carruthers, 2006, p.35). Dans cette perspective, les modules innés vont permettre de structurer l'apprentissage tout au long du développement en spécifiant la valeur innée de certaines récompenses, par exemple. Pour Carruthers (2006, p.14), l'influence des facteurs génétiques s'exprime alors à différents moments et différents endroits, tout au long du développement, par l'activation ou la désactivation de certains gènes. Dans l'ensemble, l'héritabilité repose essentiellement sur la transmission génétique des modules « typiques à l'espèce » qui ont évolué pour permettre la résolution des problèmes adaptatifs statistiquement récurrents dans l'environnement évolutionnaire (Carruthers, 2006, p.163).

Selon Carruthers (2006, p.14), la MMH fournit la seule solution possible (ou, du moins, une solution extrêmement probable) à l'évoluabilité et la robustesse d'un système fonctionnel complexe comme le cerveau. Carruthers soutient que l'évolution du cerveau est en grande partie mosaïque (2006, p.153), c'est-à-dire que les pressions sélectives vont cibler des sous-systèmes séparément modifiables. Cette modifiabilité séparée serait essentielle pour l'évoluabilité en permettant la construction ou la modification incrémentielle de sous-systèmes fonctionnellement distincts. Elle serait aussi essentielle pour permettre la robustesse puisque les perturbations ou pressions sélectives vont affecter les différents sous-systèmes

[25] Carruthers semble souscrire à la conception de l'innéisme cognitif proposée par Samuels (2002), qui signifie quelque chose de « psychologiquement primitif ». Selon cette perspective, une capacité est innée : (1) lorsque celle-ci émerge d'un développement typique pour le génotype de l'espèce (condition de normalité) et (2) lorsque celui-ci n'admet pas d'explications psychologiques sophistiquées pour son acquisition (condition de primitivité), c'est-à-dire celles qui reposent sur des processus d'apprentissage par exemple. Il faut toutefois souligner qu'il n'est pas nécessaire de résoudre la question épineuse de l'innéisme puisqu'il ne s'agit pas d'une caractéristique essentielle pour soutenir la MMH selon Carruthers (2006; p. 10).

[26] Suivant Gallistel (2000), Carruthers soutient que l'apprentissage associatif repose sur des modules d'apprentissage spécialisés dans l'estimation de fréquences de récompenses dont les opérations computationnelles ont été façonnées par l'évolution. En effet, ces modules évolués se développent en fonction de programmes développementaux génétiquement spécifiés qui génèrent des mécanismes d'apprentissage qui servent à acquérir des corpus de connaissances de plus en plus élaborées.

de façon quasi-indépendante, ce qui laisse le fonctionnement du reste du système relativement intact et donc fonctionnel. Dans le cadre de l'hypothèse de l'évolution mosaïque, la modifiabilité séparée peut être comprise dans la perspective d'une contrainte fonctionnelle sur l'évolution du cerveau. Celle-ci suggère que les composantes évoluent séparément puisque la sélection naturelle opère sur les composantes fonctionnellement interconnectées. Dans l'ensemble, nous pouvons dire que la MMH accorde une importance cruciale à la contrainte locale de modifiabilité séparée dans l'évolution de l'architecture cognitive (Carruthers, 2006, p.14). Toutefois, bien que la MMH soutient que l'évolution de multiples fonctions cognitives devrait être soutenue par la modifiabilité séparée, des contraintes de ressources et d'efficacité de traitement seraient aussi à considérer. Notamment, les contraintes de ressources semblent aller à l'encontre de la contrainte de modifiabilité séparée puisqu'elles favorisent la minimisation du volume neuronal, lequel est coûteux sur le plan énergétique (Carruthers, 2006, p.23). En conséquence, on peut s'attendre à ce qu'il y ait un compromis entre les contraintes de l'évolution mosaïque et les contraintes de ressource qui favorisent le partage de sous-composantes, pour autant que cela ne nuise pas à l'efficacité de traitement (Carruthers, 2006, p.23). En revanche, les contraintes de ressources qui tendent à minimiser la distance de connexion concernent non seulement le coût énergétique, mais aussi la vitesse de traitement, et ce, parce que la longueur des connexions neuronales est positivement corrélée à la masse (qui augmente la consommation d'énergie) et négativement corrélée à la vitesse (qui diminue avec la distance de signalement parcourue). En ce sens, le partage des sous-composantes serait avantageux pour autant qu'il n'augmente pas la distance de signalement ou le temps de traitement (Carruthers, 2006, p.25). C'est pourquoi Carruthers soutient que les contraintes de ressources et d'efficacité de traitement permettent d'ajouter de la complexité à la contrainte de modifiabilité séparée, sans pour autant remettre en cause les postulats fondamentaux de l'évolution mosaïque et de la MMH (2006, p.25).

1.3 L'hypothèse du redéploiement massif : le cerveau comme système de gestion des affordances

Dans cette section, nous introduisons les thèses complémentaires qui caractérisent l'hypothèse du redéploiement massif (MRH). Selon la MRH structuraliste et développementaliste, les capacités cognitives humaines émergent d'assemblages neuronaux transitoires qui reposent sur des propriétés fonctionnelles domaines-générales façonnées par des processus d'auto-organisation évo-dévo (Anderson, 2014; 2016). Plus précisément, trois hypothèses caractérisent la MRH : (1) Le cerveau est un système dynamique de gestion des affordances; (2) Les processus cognitifs émergent d'interactions dynamiques non-linéaires entre régions neuronales qui forment des assemblages transitoires de sous-systèmes neuronaux locaux

(TALoNS) (dans des associations structure-fonction « many-to-many »); (3) L'architecture fonctionnellement différenciée est façonnée par des pressions sélectives globales dans un processus d'auto-organisation évo-dévo. Puisque la MRH est une proposition radicale qui s'applique à l'ensemble de l'architecture cognitive, elle nous permettra de mettre en relief les thèses complémentaires qui la caractérise, soit le dynamicisme non-représentationnel, l'interactivité non-décomposable et l'adaptativité par évolution concertée. Dans les sections suivantes, nous aborderons les versions plus récentes de la MRH telles que proposées par Anderson (2014; 2016).

1.3.1 L'architecture dynamique

Comme nous l'avons précédemment abordé, la MRH souscrit au paradigme de la cognition incarnée qui soutient que la cognition s'étend dans un système cerveau-corps-environnement, la perception et l'action étant dynamiquement couplés au corps et à l'environnement dans des boucles sensorimotrices. Dotée d'un accès direct au monde, la cognition est comprise comme un processus orienté vers l'action qui repose sur la gestion des interactions dynamiquement couplées avec l'environnement. En d'autres termes, on comprend la cognition comme un processus performatif pour lequel la fonction de la perception est de guider l'action en construisant une réalité subjective centrée sur les intérêts pratiques de l'organisme (ce qu'on nomme un Umwelt; voir Uexküll, 1957 pour la notion originale). Les adeptes de la cognition incarnée associent typiquement la gestion des interactions sensorimotrices au dynamicisme, c'est-à-dire qu'on comprend le fonctionnement de la cognition comme étant analogue à celui d'un régulateur centrifuge de Watt[27] qui maintient des paramètres constants en s'ajustant continuellement aux perturbations par des boucles de rétroactions (voir Eliasmith, 1997). Dans cette perspective, les adeptes de la cognition incarnée qualifient souvent la cognition comme un processus dynamique de gestion des interactions sensorimotrices (ou des affordances[28]) qui permet le maintien des paramètres homéostatiques de l'organisme. Cependant, bien qu'il existe plusieurs façons de comprendre la gestion des interactions sensorimotrices, certains adeptes de la cognition incarnée se réfèrent souvent à quelque chose comme un processus de couplage dynamique qui implique une « connaissance implicite » des « contingences

[27] Un gouverneur centrifuge (de Watt) est un mécanisme qui permet de maintenir automatiquement la vitesse de rotation spécifique d'un axe fixé à un objet régulé (tel qu'un moteur ou une turbine) au moyen d'un senseur qui prend la forme de poids tournants; la force centrifuge des poids tournants est utilisée pour déplacer l'élément qui régule l'objet.

[28] Les affordances sont des opportunités d'action sollicitées par l'environnement, lesquelles dépendent des habiletés particulières de l'organisme (voir Gibson, 1979).

sensorimotrices » - c'est-à-dire les régularités associées à la façon dont la perception change avec l'action (O'Regan & Noë, 2001). Cependant, les contingences sensorimotrices semblent suggérer une interprétation intellectualiste qui repose sur des représentations internes, ce qui serait problématique pour certains adeptes de la cognition incarnée (voir Hutto, 2005). Une autre façon, plus pertinente celle-là, d'interpréter les interactions/contingences sensorimotrices consiste à les envisager dans la perspective d'une tendance vers une « emprise optimale » sur un « champ d'affordances » (c'est-à-dire les différentes opportunités d'action sollicitées par l'environnement) (Bruineberg & Rietveld, 2014; voir aussi Dreyfus, 2002 sur la notion d'emprise optimale).

La tendance vers une emprise optimale peut être comprise comme une conjonction entre la théorie des systèmes dynamiques et la théorie des affordances. La conjonction entre une approche dynamique de la cognition et une intentionnalité experte non-représentationnelle fut d'abord proposée par Rietveld (2008). Nous pouvons comprendre la tendance vers une emprise optimale sur le champ d'affordance comme un processus de sélection de l'action optimale en fonction du contexte. Dans cette perspective, la sélection de l'action est un processus distribué de compétition biaisée d'affordances où les trajectoires d'action alternatives compétitionnent entre elles en parallèle sur la base de biais (salience, attention, récompense, etc.) jusqu'à ce qu'une trajectoire optimale soit sélectionnée (voir Cisek, 2007, pour une description détaillée du processus de compétition biaisée d'affordances). La tendance vers une emprise optimale implique alors une « intentionnalité experte » ou motrice non-représentationnelle, laquelle est dirigée vers la réduction d'une tension dans le champ d'affordances (Bruineberg & Rietveld, 2014). En d'autres termes, la tendance vers l'emprise optimale décrit comment un agent expert, simultanément réactif aux multiples affordances sollicitées par l'environnement, peut améliorer son emprise sur des situations familières et réduire ainsi la tension ressentie. Puisqu'il y a couplage dynamique entre la perception et l'action, il n'est plus nécessaire d'employer des représentations pour guider l'action, il suffit d'acquérir l'expertise qui permet l'exploitation des régularités statistiques associée à la façon dont les sensations changent avec l'action. C'est-à-dire que cela permet des contingences sensorimotrices sans pour autant suggérer des représentations riches en contenu informationnel (Noë, 2004)[29].

[29] Cependant, cette façon d'épurer complètement toute forme de contenu représentationnel peut rendre les processus contrefactuels de haut niveau, dits « representation hungry » (tels que la planification, l'imagination, la pensée abstraite, etc.), notoirement difficiles à expliquer selon certains auteurs (Clark et Toribio, 1994). Un problème similaire est celui du « cognitive gap » qui exprime un gouffre théorique que les approches incarnées de la cognition doivent franchir pour passer des processus « online » (couplés à l'environnement) aux processus « offline » (découplés de l'environnement) sans pour autant retomber dans le cognitivisme. Dans le second

Nous pouvons comprendre la MRH d'Anderson comme une perspective radicale puisque la quasi-totalité de l'architecture serait ainsi dynamique et non-représentationnelle. C'est pourquoi le reste de la sous-section sera consacrée à présenter le dynamicisme dans cette perspective particulière. En effet, la MRH d'Anderson s'inscrit dans cette perspective orientée vers l'action (ou performative) pour laquelle la plupart des processus cognitifs reposent sur la compétition d'affordances couplées avec l'environnement. Pour être exact, les représentations ne sont pas exclues d'emblée – elles peuvent jouer un rôle périphérique dans certains processus de haut niveau, mais celles-ci reposent ultimement sur un redéploiement des ressources neuronales associées aux processus sensorimoteurs (Anderson, 2014, p.162). Dans cette perspective dynamique non-représentationnelle, l'essentiel de l'architecture cognitive humaine s'articule autour d'une variante de la compétition biaisée entre patrons d'activités distribués (ou globaux) dans des réseaux neuronaux massivement réentrants – ce qu'on comprend comme la compétition baisée d'affordances d'un point de vue psychologique (Anderson 2014; 2016). Dans le langage de la théorie des systèmes dynamiques, on comprend le cerveau comme un système dynamique qui contrôle une trajectoire continue dans un espace d'états à haute dimensionnalité représentant les degrés de liberté du corps de l'organisme[30] (Anderson, 2014, p. 217). L'évolution de la trajectoire d'états va dépendre des attracteurs sous-jacents ainsi que des inputs sensoriels qui vont perturber l'état général des réseaux du cerveau en entraînant la compétition baisée entre différents patrons d'activité distribués. Cette compétition biaisée entre patrons d'activité permettra l'établissement rapide de diverses coalition fonctionnelles en fonction de « dynamiques multi-échelles » qui reconfigurent continuellement le paysage d'attracteurs. Au cours de la compétition biaisée, les biais fonctionnels régionaux vont jouer leurs rôles causaux en renforçant ou perturbant certains patrons d'activité. Lorsqu'un patron d'activité en vient ultimement à dominer la compétition baisée, il entraîne alors d'autres changements dans le paysage

chapitre, nous soutiendrons que ce scepticisme est justifié puisque nous soutiendrons qu'une intentionnalité sans contenu (comme l'intentionnalité motrice ou experte proposée par Bruineberg et Rietveld, 2014) serait inadéquate pour décrire certains processus « top down », notamment les processus que nous appellerons « model-based » ou « dirigés vers des buts », lesquels, comme nous le verrons, reposent nécessairement sur la manipulation de représentations abstraites, découplées avec l'environnement.

[30] Dans le jargon de la théorie des systèmes dynamiques, un état est un espace abstrait qui permet de modéliser l'évolution temporelle d'un système en fonction de tous les états possibles dans lequel il peut se trouver. L'espace d'état inclut toutes les quantités pertinentes qui peuvent changer dans le système (c'est-à-dire toutes les variables pertinentes) qu'on associe à différentes dimensions. Chacune des dimensions de cet espace d'état correspond à une variable du système qu'on spécifie en assignant une valeur particulière (on assigne une position particulière au système sur chacune des dimensions), ce qui permet une spécification complète du système. La trajectoire dans cet espace d'état permet de décrire l'évolution du système en termes de flux d'états à travers le temps. Lorsqu'une trajectoire des états systémiques revient périodiquement sur elle-même pour revisiter une région de l'espace d'état, on dit alors qu'elle forme un ensemble attracteur.

d'attracteurs qui vont permettre la formation de nouvelles coalitions fonctionnelles. En d'autres termes, les patrons d'activité distribués correspondent aux boucles action-perception itératives qui répondent directement aux affordances perçues, cela sans dépendre d'une médiation représentationnelle (Anderson, 2014, p. 203). Et puisque les patrons d'activité distribués en question se chevauchent les uns les autres durant la compétition biaisée, ils ne peuvent pas se recombiner dans une structure componentielle/compositionnelle comme le suggère le LOT (Anderson, 2016, p. 6).

Pour la MRH d'Anderson, la sélection de l'action n'implique pas de médiation de représentations au cours des boucles de rétroaction. Dès lors, les boucles de rétroaction vont permettre à la perception d'ajuster directement l'action et à l'action d'ajuster directement la perception dans un couplage dynamique qui permet une coordination réciproque entre l'input et l'output. Plus précisément, Anderson soutient que la compétition entre patrons d'activité est continuellement biaisée par les inputs sensoriels, lesquels sont aussi intégrés et influencés par l'activité cérébrale en cours résultant de patrons d'activité antérieurs par des boucles de rétroaction qui ajustent directement l'output moteur (Anderson, 2016, p.6). Dans cette perspective, il ne s'agit pas de sélectionner passivement l'action optimale en fonction des sensations, mais plutôt de sélectionner (inter)activement les affordances qui guident l'action vers les sensations optimales. Ce sont alors les « contingences sensorimotrices » (c'est-à-dire les régularités associées au couplage entre action et perception) et la perception des affordances qui coopèrent pour guider l'action vers les sensations optimales. Les valeurs de contrôle et de motivation, qui seraient supposément conférées par les modules centraux dans une architecture hiérarchique, seraient alors comprises comme autant de biais fonctionnels qui se combinent dans la « rétroaction multidirectionnelle » entre processus « top-down », « bottom-up », « feedforward » et « feedback ». En ce sens, les patrons d'activité biaisés par les estimations de récompenses, la saillance, l'attention et les interactions continues avec l'environnement sont renforcés ou perturbés jusqu'à ce qu'un consensus distribué émerge et qu'une boucle action-perception soit spontanément énactée, sans intervention d'un module central de sélection de l'action. En ce qui concerne le problème de la cognition « offline », Anderson soutient que les interactions continues peuvent aussi entraîner le rappel d'une expérience perceptuelle, ce qui peut aussi biaiser la compétition de patrons d'activité. Pour Anderson, les boucles de rétroaction « offline », qui sont associées à la simulation perceptuelle, sont essentiellement les mêmes que celles qui soutiennent les boucles de rétroaction « online » et permettent de guider directement l'action. En effet, les boucles de rétroaction « online » seraient alors redéployées pour des interactions indirectes avec les traces des expériences perceptuelles emmagasinées en mémoire (Anderson, 2014, p.190). Les mêmes capacités qui permettent

d'énacter les boucles de rétroaction en relation avec des éléments présents dans l'environnement nous permettent aussi d'énacter les boucles action-perception « offline », comme si ces éléments étaient présents. Toutefois, cette conception « pragmatiste » de l'action planifiée semble tout de même impliquer des modèles internes, riches en contenu informationnel. En effet, les boucles action-perception couplées à l'environnement sont probablement redéployées lors de la planification, mais ce redéploiement réalise une fonction différente de la première. Comme nous le verrons dans le second chapitre, il est plus avantageux de concevoir la planification de l'action comme un processus computationnel qui repose sur la manipulation de représentations découplées de l'environnement.

1.3.2 L'architecture interactive

Cette conception dynamique non-représentationnelle de l'architecture cognitive entraîne un certain nombre de conséquences théoriques. En effet, la cognition est comprise comme un processus complexe qui repose sur la gestion d'interactions sensorimotrices riches en contenu informationnel. Puisque cette conception de la cognition s'appuie sur l'intégration des ressources internes et externes dans un système cerveau-corps-environnement, on suppose qu'une externalisation de l'information dans le corps et l'environnement permet de réduire la complexité de traitement. Dans cette perspective, l'architecture cognitive est comprise comme un système dynamique non-décomposable dont les interactions non-linéaires multiplicatives[31] entre composantes sur plusieurs échelles temporelles sont déterminées par des dynamiques globales entre cerveau, corps et environnement. Contrairement aux systèmes quasi-décomposables, c'est l'interactivité qui domine dans un système non-décomposable (dominance interactive), c'est-à-dire que les interactions entre composantes sont plus importantes que les interactions au sein des composantes (Van Orden, Holden et Turvey, 2003). On considère qu'un système dynamique complexe est non-décomposable quand des interactions multiplicatives entre processus interdépendants génèrent des propriétés émergentes (Van Orden, Holden et Turvey, 2003). Plus spécifiquement, nous pouvons distinguer trois propriétés importantes pour la non-décomposabilité (voir Richardson & Chemero, 2014). Premièrement, l'interdépendance entre les composantes est réalisée sur plusieurs échelles temporelles dans le cerveau, le corps et l'environnement. En d'autres termes, les interactions dynamiques entre le cerveau, le corps et l'environnement sont profondément entrelacées, elles ne peuvent pas être isolées dans des composantes indépendantes. Deuxièmement, les interactions non-linéaires

[31] Un système dynamique non-linéaire est un système dont les changements dans les outputs ne sont pas proportionnels aux changements dans l'input. Ce faisant, les variables des systèmes dynamiques non-linéaires sont multiplicatives, multidimensionnelles et difficiles à quantifier.

multiplicatives entre composantes génèrent des structures émergentes qui sont irréductibles aux composantes elles-mêmes. Troisièmement, les structures émergent d'un processus d'auto-organisation qui intègre les composantes dans des ensembles coordonnés (sans l'intervention d'un contrôle central), ce qu'on appelle des « soft-assemblies ».

Dans une perspective dynamique, la non-décomposabilité est typiquement associée à la notion de « soft-assembly ». La notion de « soft-assembly » réfère aux composantes d'un système qui s'auto-organise pour générer une fonction d'une façon temporaire et sensible au contexte (Kello & Van Orden, 2009). Dans cette conception, l'accent est mis sur la synthèse compositionnelle flexible, l'incorporation dynamique de ressources externes et la plasticité adaptative du système (Schiavio et Kimmel, 2021). Une synthèse compositionnelle flexible et adaptative est essentielle pour permettre une réorganisation dynamique des composantes en fonction des exigences particulières de la tâche ou du contexte. Dans les dynamiques des réseaux neuronaux, les fonctions sont « soft-assembled » dans des patrons d'activité en équilibre transitoire ou « métastable » se trouvant au carrefour entre indépendance (ou ségrégation locale) et dépendance (ou intégration globale) (Kello & Van Orden, 2009). On distingue alors les fonctions « soft-assembled » des fonctions hard-molded[32]/hard-wired qui reposent sur des composantes rigides qui réalisent des fonctions spécifiques (Anderson, Richardson et Chemero, 2012). En effet, les composantes des fonctions « soft-assembled » peuvent être remplacées les unes par les autres sans grandes conséquences puisque les fonctions émergent principalement des interactions entre composantes, et non des composantes elles-mêmes. Par exemple, nous pouvons intersubstituer les composantes qui forment un essaim d'insectes, un banc de poissons ou une envolée d'oiseaux sans engendrer des conséquences fonctionnelles importantes. Dans les fonctions hard-molded/hard-wired, les composantes ne peuvent pas être remplacées par d'autres composantes sans conséquences puisque chacune des composantes réalisent des fonctions précises. Par exemple, nous ne pouvons pas intersubstituer les composantes d'un ordinateur ou d'une voiture sans engendrer des conséquences fonctionnelles importantes.

Les « soft-assemblies » permettent aussi d'incorporer les ressources externes au système puisque les contributions fonctionnelles des composantes du système sont indéterminées dans les dynamiques à dominance interactive. En effet, les « soft-assemblies » sont des coalitions temporaires qui émergent

[32] Dans un système dynamique à dominance componentielle, les composantes « hard-molded » réalisent des fonctions prédéterminées et « insensibles au contexte » en raison de la prédominance de leurs dynamiques intrinsèques.

d'interactions dynamiques multiéchelles entre composantes sur plusieurs échelles temporelles et spatiales, ce qui coupent à travers les frontières entre cerveau, corps et environnement. En ce sens, on se réfère parfois au terme de synergies pour parler des soft-assemblies puisqu'elles forment des regroupements d'éléments structurels (neurones, muscles, membres, etc.) qui sont temporairement contraints pour agir ensemble dans un tout fonctionnel cohésif, toujours sans contrôle central (Kelso, 2009). Puisqu'elles sont maintenues ou modifiées spontanément en fonction de la tâche et de l'environnement dans lequel les processus évoluent à travers le temps, les fonctions des synergies sont distribuées plutôt que localisées dans des composantes. Et puisque les fonctions émergent des interactions non-linéaires (multiplicatives) entre composantes qui peuvent avoir plusieurs fonctions différentes selon les exigences de la tâche, les « soft-assemblies » confèrent une plasticité adaptative significative au système. En effet, les « soft-assemblies » mettent l'accent sur la plasticité des composantes du système puisque la fonction émerge des interactions dynamiques entre composantes, lesquelles peuvent donc acquérir des nouvelles fonctions tout en conservant l'ancienne. Cette particularité permet de réaliser un nouveau type de plasticité neuronale, la réutilisation neuronale, où les éléments neuronaux qui remplissent une fonction peuvent être réutilisés, redéployés, recyclés ou « exaptés » pendant l'évolution et le développement pour servir d'autres fonctions, généralement sans entraîner la perte de la fonction d'origine (Anderson, 2014).

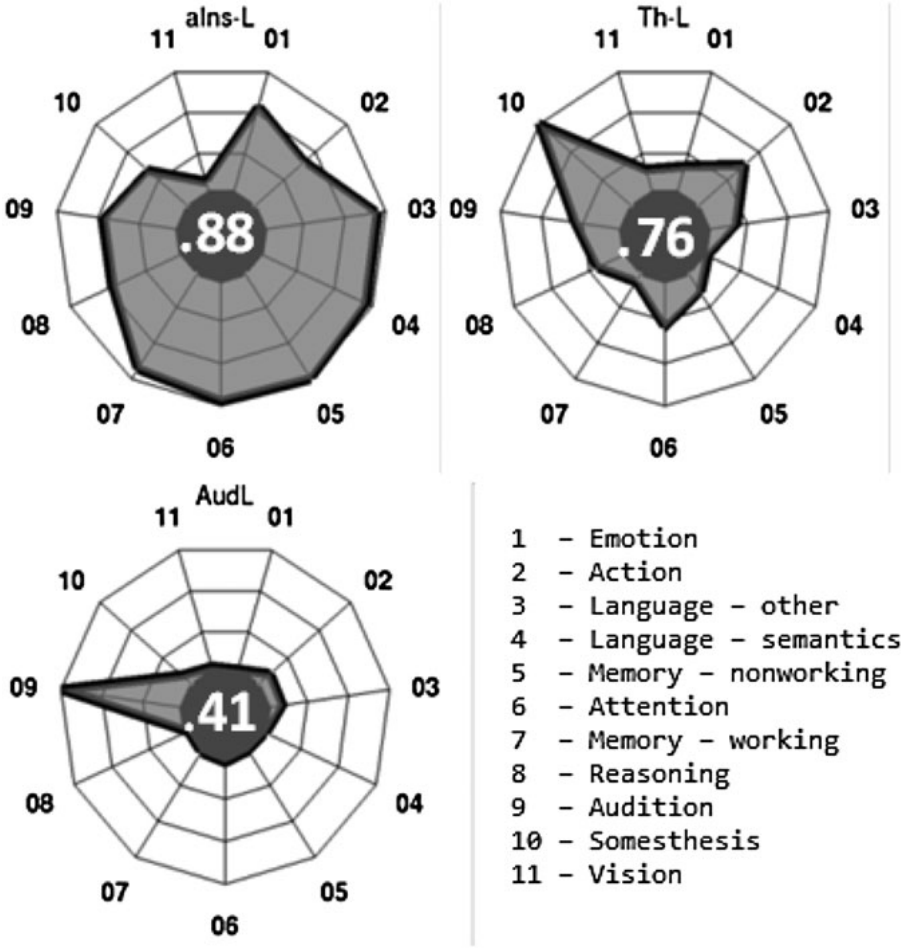
On peut envisager la MRH comme une perspective radicale qui pousse cette conception interactive non-décomposable de l'architecture cognitive jusqu'à ses limites conceptuelles. C'est pourquoi nous consacrerons le reste de cette sous-section à présenter l'interactivité dans cette perspective particulière. Dans l'ensemble, la MRH soutient que la quasi-totalité des processus cognitifs émergent d'interactions distribuées entre régions neuronales fonctionnellement différenciées dans des assemblages transitoires de sous-systèmes neuronaux locaux (Transiently Assembled Local Neuronal Subsystems ou TALoNS, Anderson 2014, p. 94; voir Buzsaki 2019 pour une notion similaire). Les TALoNS d'Anderson sont profondément ancrés dans une conception biologique qui repose sur la notion de « soft-assemblies » (Anderson, Richardson et Chemero, 2012). En ce sens, les TALoNS sont des « soft-assemblies » qui réalisent des fonctions temporaires, mais reproductibles, lesquelles fonctions sont réalisées par les interactions entre les composantes plutôt que par les composantes elles-mêmes. En effet, la sélectivité fonctionnelle des TALoNS découle de la manière dont les propriétés des composantes (multi)fonctionnelles locales, ou facteurs NRP dispositionnels (voir Anderson, 2014, p. 145), interagissent avec d'autres structures fonctionnelles globales qui les contraignent pour former des rétroactions

multidirectionnelles. Ces interactions dynamiques non-linéaires entre contraintes « top-down » et constituants « bottom-up » vont permettre l'établissement dynamique de partenariats fonctionnels sensibles au contexte qui peuvent traverser la frontière entre l'organisme et l'environnement pour incorporer des composantes extra-neuronales (notamment, les interactions avec d'autres individus par le biais d'artefacts symboliques). Dans cette perspective, les composantes neuronales locales peuvent participer à plusieurs coalitions fonctionnelles différentes, lesquelles sont non-décomposables puisqu'elles vont se superposer significativement durant la compétition biaisée d'affordances. Par conséquent, on peut comprendre les TALoNS dans la perspective d'une coordination synergique puisque les composantes neuronales se coordonnent avec celles du corps et de l'environnement sur plusieurs échelles temporelles, ce qui permet un contrôle décentralisé et (possiblement) hétérarchique.

Selon Anderson (2014, 2016), la contribution des régions fonctionnellement différenciées est analysable en identifiant leurs propriétés dispositionnelles « domaine-générale » (les facteurs neuroscientifiquement pertinents ou facteurs NRP) dans des empreintes fonctionnelles multidimensionnelles (association structure-fonction « many-to-many »). Puisque les TALoNS sont des sous-systèmes non-décomposables qui dépendent d'interactions dynamiques non-linéaires de haut niveau, l'analyse par décomposition et localisation mécaniste préconisée par Bechtel et Richardson (2010) doit selon Anderson (2014, p. 117) être abandonnée au profit d'empreintes fonctionnelles multidimensionnelles qui vont permettre de capturer les dispositions fonctionnelles sous-jacentes (voir figure 4). L'analyse des empreintes fonctionnelles multidimensionnelles permettrait de révéler des ensembles communs de facteurs NRP primitifs qui, collectivement, vont caractériser les « personnalités neuronales » (ou dispositions fonctionnelles) de régions individuelles (Anderson 2014; p. 137). Dans cette perspective, les régions neuronales individuelles vont seulement différer dans leurs charges (ou « loading », dans le sens que lui donne l'analyse factorielle au cœur de ces analyses) sur ces ensembles de facteurs NRP primitifs plutôt que sur les opérations de leurs composantes. De manière spéculative, Anderson suppose que ces facteurs PNR peuvent être compris comme des biais régionaux qui aident à gérer les valeurs de certains aspects des interactions de l'organisme avec l'environnement, bien que leur nature exacte reste encore à être identifiée selon Anderson (2014 p. 146). Quelle qu'elle soit, la nature exacte des facteurs NRP qui sont pertinents pour comprendre la cognition humaine, elle ne coïncide probablement pas selon Anderson avec la taxonomie actuelle des fonctions cognitives, laquelle est majoritairement héritée de la psychologie cognitive (qui nous provient de la psychologie du 19^{e} siècle, notamment sous l'influence de James 1890). Cela impliquerait une révision bottom-up de l'ontologie cognitive qui dépend d'associations entre patrons

interactifs d'activation et diverses caractéristiques environnementales (Anderson 2014 p. 114). Étant donné que la sélectivité fonctionnelle des TALoNS émergent d'interactions dynamiques globales (sur plusieurs niveaux) qui contraignent les contributions fonctionnelles locales des régions neuronales et non l'inverse, on doit substituer la décomposition mécaniste par la notion de contraintes habilitantes (ou « enabling constraints ») selon Anderson (2015; voir section 1.1). Ce faisant, il est possible de capturer les interactions entre les composantes locales et les interactions globales qui les contraignent (plutôt que des interactions linéaires additives entre composantes stables et semi-indépendantes sur le même niveau).

Figure 4 : Empreintes fonctionnelles multidimensionnelles chez Anderson



Empreintes fonctionnelles multidimensionnelles représentant le degré d'activité pour 11 catégories de tâches dans le thalamus (Th-L), l'insula antérieur gauche (ains-L) et le cortex auditif gauche (AudL). Nous avons donc ici une représentation de la « personnalité » de ces trois régions cérébrales (tirée de Anderson, 2016, p. 3).

1.3.3 L'architecture adaptative

Il nous reste maintenant une dernière thèse importante à aborder dans notre survol des thèses principales de cette version du paradigme de la cognition incarnée. Dans la mesure où la cognition est un processus performatif qui repose sur des interactions (physiques ou sociales) incarnées riches en contenu informationnel, les adeptes de la cognition incarnée vont pouvoir répondre au problème de sous-détermination de l'expérience (voir section 1.1.3). Selon l'argument de la richesse du stimulus, développé par Sterelny (2003) en réponse à l'argument de la pauvreté du stimulus de Chomsky, l'information en provenance de l'environnement développemental serait suffisamment riche pour permettre l'apprentissage des capacités cognitives typiquement humaines. Selon certaines interprétations de l'argument, cela relève du fait qu'on postule que les interactions incarnées permettent d'extraire suffisamment d'information pour supplémenter l'information génétique, que ce soit l'information de l'environnement développemental et/ou l'information culturellement héritée. Sur la base de cet argument, les adeptes de la cognition incarnée vont se fonder sur les interactions incarnées pour expliquer des façons alternatives de concevoir l'acquisition du langage (Cuffari, Di Paolo & Jaegher, 2015; Hutto, 2007), la théorie de l'esprit (Gallagher, 2001; Hutto, 2007), etc. Selon cette perspective, l'acquisition du langage ou la théorie de l'esprit sont des pratiques socioculturelles (ou pratiques narratives, par exemple; voir Gallagher & Hutto, 2008) interactives acquises au cours du développement. Ces pratiques socioculturelles interactives reposent sur une construction de niche qui permet un couplage réciproque/bidirectionnel avec l'environnement de sélection dans la perspective d'une coévolution gène-culture (voir Laland et al., 2000). Puisque la construction de niche peut entraîner des changements rapides dans l'environnement de sélection, elle aurait tendance à favoriser l'évolution d'une plasticité adaptative significative (Sterelny, 2003). Dans cette perspective, on envisage les pratiques socioculturelles interactives dans la perspective d'une synthèse évolutionnaire étendue qui permet d'assimiler divers mécanismes de transmission de l'information tels que la construction de niche (culturelle), la plasticité, théories évo-dévo, les systèmes développementaux, la transmission épigénétique, etc.

Pour les partisans de la synthèse évolutionnaire étendue[33], la sélection naturelle n'est pas la seule force capable d'expliquer l'évolution de systèmes fonctionnels complexes comme le cerveau (Anderson & Finlay, 2014). Dans cette perspective, l'architecture cognitive humaine résulte non seulement d'un processus de

[33] Contrairement à l'approche néo-darwinienne (ou synthèse moderne), la synthèse évolutionnaire étendue est un regroupement hétérogène d'approches qui entretiennent parfois des tensions théoriques importantes. Dans ce qui suit, nous présenterons les approches qui s'accordent avec la proposition d'Anderson (2014; 2016).

sélection naturelle, mais aussi d'un processus d'auto-organisation qui permet l'assemblage reproductible des ressources développementales récurrentes qui émergent d'interactions dynamiques avec l'environnement (socioculturel) (Anderson & Finlay, 2014). Cette conception de l'évolution des systèmes fonctionnels complexes s'articule principalement autour d'une contrainte de (réutilisation des) ressources (Cherniak et al. 2004; Anderson, 2010), ce qui implique la non-décomposabilité puisque les mêmes ressources sont impliquées dans différents assemblages. Dans cette perspective, l'évoluabilité des systèmes fonctionnels complexes implique des pressions sélectives sur les processus développementaux, ce qui permet une plasticité adaptative significative en réponse aux changements environnementaux. En biologie évolutionnaire, cette conception de l'évoluabilité s'articulant autour des contraintes développementales correspond à l'hypothèse de l'évolution concertée, laquelle implique des pressions sélectives globales qui agissent sur l'ensemble du cerveau (Finley & Darlington, 1995). Dans cette perspective, si le volume (ou toutes autres propriétés) d'un organe évolue de concert avec le volume des autres organes, son évolution peut être considérée comme étant concertée. L'évolution concertée du cerveau bénéficierait d'un certain degré de support empirique, notamment en ce qui concerne l'existence de contraintes développementales sur les changements de volume (Finlay & Darlington, 1995; Whiting & Barton, 2003). En résumé, l'hypothèse de l'évolution concertée repose essentiellement sur des contraintes développementales (globales) qui suggèrent que les composantes tendent à évoluer ensemble puisque les pressions sélectives agissent sur des processus développementaux qui affectent toutes les composantes des réseaux du cerveau de façon concertée (Charvet and Finlay, 2012).

On peut comprendre la MRH comme une perspective radicale puisque la quasi-totalité de l'architecture cognitive repose sur une différenciation fonctionnelle qui peut s'adapter rapidement à l'environnement changeant (contrairement aux modules qui sont des adaptations à l'environnement évolutionnaire). La MRH implique une part importante de plasticité adaptative qui repose, selon Anderson (2014, p.49), sur un processus de différenciation et de recherche interactive (interactive differentiation and search, IDS), lequel permet d'ajuster les différentes charges régionales sur les facteurs NRP « domaine-généraux » en fonction d'interactions interrégionales dépendantes de l'activité. L'IDS permet d'expliquer comment un développement fonctionnel flexible peut émerger d'une différenciation interactive locale des biais fonctionnels complexes et d'une recherche neuronale globale qui permet l'évaluation rapide de multiples partenariats neuronaux pour identifier les options fonctionnelles adéquates qui seront subséquemment réalisées dans des TALoNS (Anderson, 2014, p.53). Selon l'IDS, l'apprentissage consiste à trouver et consolider les partenariats neuronaux adéquats pour supporter l'acquisition des comportements, c'est-à-

dire trouver les régions neuronales avec les biais fonctionnels adéquats pour réaliser la tâche. Dans cette perspective, les biais fonctionnels sont façonnés par l'apprentissage et l'expérience dès le tout début du développement (Anderson, 2016, p.5). Pour Anderson et Finlay (2014, p.23), l'influence des facteurs génétiques s'exprime surtout dans des projections neuronales hautement stéréotypées des afférences sensorielles. Cela serait suffisant, selon eux, pour permettre l'héritabilité des biais fonctionnels qui forment des assemblages neuronaux typiques à l'espèce, notamment quand on considère que les propriétés statistiques de l'environnement sont largement conservées entre les générations en raison de différents degrés de construction de niche (Anderson & Finlay, 2014, p.23).

Selon Anderson, la MRH permet de fournir une solution alternative au problème de l'évoluabilité et de la robustesse de l'architecture cognitive qui n'implique pas d'évolution mosaïque. Pour Anderson (Anderson & Finlay, 2014, p.20), l'évolution du cerveau est concertée de manière importante, les pressions sélectives globales agissant globalement en ciblant des processus développementaux pour permettre la disponibilité robuste d'une vaste gamme de biais fonctionnels dans tout le cerveau. La réutilisation neuronale permet l'évoluabilité en permettant des biais fonctionnels flexibles qui peuvent établir des partenariats fonctionnels inédits pour supporter de nouvelles capacités comportementales. La réutilisation neuronale permet aussi la robustesse aux perturbations en provenance de l'environnement puisque les fonctions distribuées reposent sur des biais fonctionnels qui peuvent s'assembler dans différentes configurations, ce qui permet la formation de nouveaux partenariats fonctionnels pour compenser les perturbations/répondre aux pressions sélectives. Dans le cadre de l'hypothèse de l'évolution concertée, on peut comprendre la réutilisation neuronale dans la perspective d'une contrainte développementale qui suggère que les composantes du cerveau évoluent ensemble puisque la sélection naturelle opère sur les processus développementaux qui affectent la croissance de toutes les composantes de façon concertée. En ce sens, la MRH propose une emphase sur la contrainte globale de (réutilisation des) ressources dans l'architecture cognitive (Anderson, 2010, p. 248). En effet, la MRH semble compatible avec la théorie de l'optimisation de la connectivité globale pour les raisons suivantes (Cherniak et al., 2004) : Premièrement, la longueur des connexions n'est pas la seule contrainte, mais la masse neuronale totale requise pour réaliser les fonctions doit être maintenue au minimum, et la MRH serait une bonne façon d'y parvenir. Deuxièmement, il faut garder à l'esprit que la théorie de l'optimisation prédit l'optimisation globale des composantes, ce qui est compatible avec des sous-ensembles de composantes avec une connectivité sous-optimale. Troisièmement, il n'y a aucune raison de s'attendre à ce que tous les sous-ensembles de composantes soient sous-optimaux de la même façon; l'optimalité globale est compatible avec des

différences dans l'optimalité des différents sous-ensembles de composantes. Quatrièmement, lorsqu'il y a une différente dans l'optimalité dans des sous-ensembles de composantes, la MRH prédit que ces différences suivraient l'âge évolutif de la fonction supportée par ces composantes, c'est-à-dire que les sous-ensembles de composantes qui supportent des fonctions plus récentes (le langage, par exemple) auraient tendance à être configurées de façon moins optimale que les fonctions plus anciennes, qui sont généralement moins distribuées.

1.4 Conclusion de chapitre

Dans ce chapitre, nous avons proposé d'envisager l'architecture cognitive dans la perspective d'un pluralisme explicatif intégratif qui distingue entre explications fonctionnelles et structurelles en philosophie de la biologie (Linde Medina, 2010). D'une part, les explications fonctionnalistes sélectionnistes stipulent que des contraintes fonctionnelles vont favoriser l'émergence d'une organisation componentielle puisque les changements évolutionnaires vont cibler des parties fonctionnelles de façon mosaique (Barton & Harvey, 2000). Comme nous l'avons vu, les trois thèses principales de la MMH souscrivent à une organisation à « dominance componentielle » dans laquelles les propriétés fonctionnelles des parties déterminent les propriétés fonctionnelles du tout (Van Orden, Holden et Turvey, 2003). En effet, notre survol des thèses de l'architecture cognitive computationnelle, modulaire et adaptée démontre qu'elles reposent sur des composantes locales dont les fonctions résultent essentiellement de la sélection naturelle dans l'environnement évolutionnaire. D'autre part, les explications structuralistes développementalistes stipulent que des contraintes développementales vont favoriser l'émergence d'une organisation non-componentielle puisque les changements évolutionnaires vont cibler des processus développementaux qui affectent l'organisme de façon concertée (Finlay & Darlington, 1995). Comme nous l'avons vu, les trois thèses principales de la MRH souscrivent à une organisation à « dominance interactive » dans laquelle les propriétés fonctionnelles du tout déterminent les propriétés fonctionnelles des parties (Van Orden, Holden et Turvey, 2003). En effet, notre survol des thèses de l'architecture cognitive dynamique, interactive et adaptative démontrent qu'elles reposent sur des processus globaux dont les fonctions résultent de l'auto-organisation dans l'environnement développemental. Dans le prochain chapitre, nous proposons d'envisager les explications fonctionnelles et structurelles sous le FEP, ce qui nous permettra de comprendre en quoi les thèses de la MMH et de la MRH sont ultimement complémentaires dans une explication adéquate de l'architecture cognitive.

CHAPITRE 2

LES PARADIGMES ANTAGONISTES SOUS LE PRINCIPE D'ÉNERGIE LIBRE: L'ARCHITECTURE COGNITIVE MODULAIRE ET INTERACTIVE

Dans ce chapitre, nous proposons d'aborder les thèses qui caractérisent l'hypothèse de la modularité massive (MMH) et l'hypothèse du redéploiement massif (MRH) dans la perspective d'un pluralisme explicatif intégratif sous le principe d'énergie libre (FEP). Après avoir présenté le FEP dans la section 2.1, nous proposons d'envisager l'architecture cognitive dans la perspective de l'esprit hiérarchiquement mécaniste (HMM), une théorie interdisciplinaire du cerveau incarné qui permet d'intégrer les explications fonctionnalistes et structuralistes. Tel que nous l'abordons dans la section 2.2, la HMM soutient que le cerveau est un système adaptatif qui minimise activement l'entropie de ses états internes par des cycles action-perception générés par des dynamiques neuronales hiérarchiques entre mécanismes neurocognitifs différentiellement ségrégués et intégrés (Badcock et al., 2019). De plus, il s'avère que la HMM est dérivée d'une théorie des systèmes évolutionnaires qui explique les aspects fonctionnalistes (parties au tout) et structuralistes (tout aux parties) de l'architecture cognitive en termes d'interaction entre la sélection et l'auto-organisation se produisant aux différents niveaux d'analyse biologique, tels que formulés par Tinbergen (adaptation, phylogénie, ontogénie, mécanisme) (Badcock, 2012; Badcock et al., 2019). Prenant en compte ces éléments, nous proposons ensuite d'envisager les thèses principales de la MMH et de la MRH comme décrivant des aspects complémentaires de l'architecture cognitive : (2.2.1) Le cerveau est un système dynamique et computationnel qui permet la manipulation de représentations lors des processus délibérés ou « dirigés vers des buts » et la gestion des affordances lors des processus automatiques ou « habituels/pavloviens »; (2.2.2) Les processus cognitifs résultent d'une compromis entre la ségrégation locale modulaire, médiée par des connections denses de courtes portées et l'intégration globale intermodulaire dans des hubs « rich-club », médiée par des connections éparses de longue portée; (2.2.3) L'architecture fonctionnellement hétérogène est façonnée par un compromis entre des pressions sélectives locales pour des modules fonctionnellement spécialisés et des pressions sélectives globales pour des hubs « rich-club » multifonctionnels. Dans l'ensemble, ce chapitre a pour objet de mettre en lumière les relations ultimement complémentaires qu'entretiennent les thèses qui caractérisent les architectures cognitives de la MMH et de la MRH, soit la relation entre computationnalisme représentationnel et dynamicisme non-représentationnel, la relation entre modularité quasi-

décomposable et interactivité non-décomposable et la relation entre adaptation mosaïque et adaptativité concertée.

2.1 Une théorie globale du cerveau sous le principe d'énergie libre

Dans cette section, nous introduisons le principe d'énergie libre (FEP) en tant que théorie globale qui unifie plusieurs théories du cerveau (théorie du contrôle optimal, théorie de l'apprentissage par renforcement, théorie de l'information, darwinisme neuronal, etc., voir Friston (2010) pour une liste plus complète) sous le thème de l'optimisation. Nous soutiendrons que le formalisme du FEP permet d'accommoder à la fois les thèses complémentaires qui caractérisent le cognitivisme, représentée ici par la MMH, et celles qui caractérisent la cognition incarnée, représentée ici par la MRH. Dans l'ensemble, nous pouvons comprendre le FEP comme une théorie qui subsume les approches bayésiennes de la cognition, lesquelles décrivent le cerveau comme une machine à inférence. Comme pour l'hypothèse du cerveau bayésien (Knill & Pouget, 2004), le FEP soutient que le cerveau est une sorte de système statistique qui effectue des inférences bayésiennes dans des conditions d'incertitude (Friston, 2010). Plus précisément, les approches bayésiennes soutiennent que le cerveau implémente (ou incarne, dans le cas échéant) un modèle génératif hiérarchique qui génère des prédictions à propos des causes cachées (non-observables et devant être inférées) des sensations. Techniquement, les modèles génératifs sont des modèles statistiques qui calculent la distribution postérieure (approximative) de la manière dont les sensations sont générées à partir d'une distribution de croyance antérieure sur les états cachés du monde et d'une fonction de vraisemblance de ces états cachés, compte tenu des observations (ou évidence) sensorielles. Dans le jargon du codage prédictif (Rao & Ballard, 1999; Hohwy, 2013; Clark, 2013) – un modèle important qui explique comment le cerveau effectue des inférences bayésiennes – les modèles génératifs hiérarchiques sont des modèles statistiques qui minimisent les erreurs de prédictions en réduisant l'écart entre les sensations entrantes provenant de l'environnement et les prédictions descendantes provenant des modèles génératifs. En termes d'implémentation neurobiologique dans des microcircuits corticaux canoniques[34] (Bastos et al., 2012), on suppose que les prédictions descendantes, encodées par des cellules pyramidales profondes (unités représentationnelles), sont transmises au niveau inférieur pour supprimer les erreurs, et que les erreurs de prédiction ascendantes, encodées par des cellules pyramidales superficielles (unités d'erreurs), sont transmises au niveau supérieur pour réviser les prédictions, ce qui

[34] Les microcircuits canoniques sont des modèles des colonnes corticales dont les connexions spécifiques entre couches laminaires sont supposées sous-tendre les computations requises pour la cognition complexe (Bastos et al., 2012).

permet la minimisation des erreurs de prédiction. Fait important, les erreurs de prédictions sont ajustées par la précision anticipée de la prédiction (soit l'inverse de sa variance), laquelle permet de déterminer l'influence relative des erreurs de prédiction par rapport aux prédictions. On suppose que cet ajustement de la précision est médié par la neuromodulation et serait associée aux processus de sélection attentionnelle et d'atténuation sensorielle (Feldman et Friston, 2010).

Incorporant le codage prédictif dans son formalisme, le FEP est une théorie plus générale qui soutient que les systèmes biologiques (comme le cerveau) doivent impérativement minimiser une quantité informationnelle nommée énergie libre variationnelle s'ils espèrent survivre (Friston, 2010). Techniquement, l'énergie libre variationnelle constitue une « limite supérieure » sur l'entropie informationnelle d'un modèle génératif engendré par les états internes d'un système biologique. Dans cette perspective, l'entropie réfère à la moyenne (à long terme) de la « surprise », soit la probabilité (logarithmique négative) de rencontrer des états sensoriels en fonction d'un modèle génératif. En des termes plus simples, les systèmes biologiques limitent continuellement les états sensoriels disponibles par des cycles action-perception qui guident l'organisme vers des états adaptatifs avec peu de « surprise ». Ainsi, un système biologique parvient à survivre en se limitant à un régime d'états attractifs de faible entropie (ou d'états adaptatifs) qui sont prescrits par son phénotype. Pour prendre un exemple commun dans la littérature, un poisson minimise l'énergie libre en évitant l'état de « surprise » de se trouver hors de l'eau. Nous pouvons comprendre cette tendance à minimiser la « surprise » comme une conséquence de la sélection naturelle qui favorise les systèmes autoorganisés qui sont capables de survivre en évitant des changements d'états délétères (comme se trouver hors de l'eau pour un poisson). Pour y parvenir, les systèmes biologiques utilisent des modèles génératifs hiérarchiques encodés dans leurs états internes pour générer des prédictions qu'ils réalisent par la suite par des séquences d'actions (nommées « action policies ») dans une sorte de prophétie autoréalisatrice (i.e., ils produisent le comportement prédit par leur modèles génératifs). En ce sens, le FEP subsume le codage prédictif puisque les systèmes biologiques avec un cerveau minimisent l'énergie libre en minimisant les erreurs de prédiction (ajustées par la précision) (Friston, 2010). La minimisation de la « surprise » peut s'effectuer de deux façons complémentaires : (1) en changeant les prédictions par la modification des états internes – ce qu'on nomme l'inférence perceptuelle (et l'apprentissage sur une échelle temporelle plus longue); et (2) en changeant l'environnement de façon qu'il corresponde aux prédictions – ce qu'on nomme l'inférence active. Ici, la perception et l'action entretiennent une relation synergique réciproque dans des boucles

action-perception qui permettent de maintenir l'homéostasie de l'organisme et d'optimiser les modèles génératifs du monde.

Nous pouvons maintenant introduire une distinction importante entre l'inférence active simple qui permet de minimiser l'énergie libre variationnelle dans le présent par des séquences d'actions directes et l'inférence active adaptative qui permet de minimiser l'énergie libre attendue (ou incertitude) dans le futur par des séquences d'actions planifiées (Kirchhoff et al., 2018). En effet, l'inférence active conçoit les systèmes biologiques comme étant comme eux-mêmes des modèles, mais ce ne sont pas tous les systèmes biologique/modèles qui sont eux-mêmes capables d'utiliser des modèles pour sélectionner adaptativement l'action (Saffron, 2020). Selon cette conception, certains systèmes biologiques ne sont simplement que des modèles en interaction avec l'environnement qu'ils modélisent. Toutefois, certains systèmes biologiques/modèles possèdent des sous-systèmes capables d'utiliser des modèles génératifs hiérarchiques avec une profondeur temporelle, c'est-à-dire des modèles qui ont une grande portée temporelle dans le passé et le futur, et une richesse contrefactuelle, c'est-à-dire des modèles qui utilisent une grande variété d'hypothèses alternatives pour expliquer les sensations (Friston, 2018). La présence des sous-systèmes en question est en fait la raison d'être des cerveaux : ils permettent aux systèmes biologiques/modèles d'échanger avec l'environnement en modélisant non seulement l'environnement présent, mais aussi les environnements possibles en fonction de la planification contrefactuelle d'actions futures par la minimisation de l'énergie libre attendue (Parr & Friston, 2017). En fait, les cerveaux sont des modèles génératifs profonds qui possèdent des capacités prédictives particulières qui permettent non seulement de modéliser des « dynamiques sensorimotrices » transitoires aux niveaux inférieurs (affordances), mais aussi de modéliser, aux niveaux supérieurs de la hiérarchie, ce que Pezzulo et ses collègues nomment des « narratifs sémantiques » temporellement étendus[35] (Pezzulo, Rigoli et Friston, 2015). Plus intéressant encore, les modèles génératifs profonds offrent aussi la possibilité d'un découplage fonctionnel entre la modélisation des « narratifs sémantiques » aux niveaux supérieurs et les « dynamiques sensorimotrices » aux niveaux inférieurs, ce qui est essentiel pour permettre la planification contrefactuelle de l'action par minimisation de l'énergie libre attendue (Pezzulo, Rigoli et Friston, 2018).

[35] Selon cette proposition, les narratifs sémantiques sont définis dans une perspective déflationniste comme des « séquences générées à l'interne » qui sont impliquées dans une planification « dirigée vers des buts » (Pezzulo et al., 2014). Ces séquences hippocampiques prennent en charge la navigation par cartes mentales « online » et la répétition « offline » de l'expérience. En effet, ceux-ci sont qualifiés de narratifs sémantiques car ils impliquent l'intégration des représentations épisodiques du « quand » situées dans le cortex paralimbique avec des représentations de contenu du « quoi » situées dans le cortex associatif (Friston & Buzsaki, 2016).

En effet, les modèles génératifs profonds qui permettent la planification contrefactuelle de l'action supporte une fonction particulière; l'utilisation vicariante ou découplée des « séquences générées à l'interne » pour modéliser les conséquences futures de l'action, ce qui aurait conféré un avantage évolutionnaire significatif et favorisé la construction de niche (socioculturelle) chez l'humain (voir Sims & Pezzulo, 2021).

Comme nous l'avons mentionné au tout début de cette section, le FEP est une théorie unifiée du cerveau qui incorpore l'apprentissage par renforcement – selon lequel un agent apprend des séquences d'actions pour maximiser la somme des récompenses attendues (Friston, 2010). Ainsi, nous pouvons aussi comprendre la distinction entre l'inférence active superficielle et l'inférence active profonde dans la perspective des modèles computationnels du renforcement. Pour ce faire, nous devons spécifier, en suivant Friston et ses collègues, que l'énergie libre attendue est une quantité informationnelle décomposable en aspect épistémique (c'est-à-dire la valeur intrinsèque), qui minimise l'incertitude, et aspect pragmatique (c'est-à-dire valeur extrinsèque), qui maximise la récompense (Friston et al., 2016). En ce sens, l'inférence active profonde offre une solution au dilemme exploration-exploitation (un problème notoire en théorie du renforcement) puisque les agents agissent pour minimiser une seule quantité : l'énergie libre attendue. En effet, les agents minimisent la valeur épistémique lorsqu'ils explorent l'environnement (par exemple, lorsqu'un faucon survole un champ en quête d'un lièvre), ce qui permet de minimiser la valeur pragmatique lorsqu'ils exploitent les récompenses révélées sans ambiguïté (par exemple, lorsque le faucon s'abat sur le lièvre qu'il vient de localiser). En résumé, les actions épistémiques minimisent l'ambiguïté (exploration) pour permettre aux actions pragmatiques de maximiser la récompense (exploitation), ce qui permet l'émergence progressive d'habitudes lorsque le contexte demeure sans ambiguïté (Friston et al., 2016). Dans ce modèle hiérarchique, la présence d'incertitude contextuelle engage des inférences profondes, lesquelles sont associées au contrôle de l'action « dirigé vers des buts » (ou délibérée) et impliquent le contenu de la mémoire déclarative (Pezzulo et al., 2016). Inversement, une absence relative d'incertitude engage des inférences superficielles, lesquelles sont associées au contrôle de l'action « habituelle » (ou automatique) et impliquent le contenu de la mémoire procédurale (Pezzulo et al., 2016). En termes de neurophysiologie, le contrôle « dirigé vers des buts » engage généralement les boucles thalamocorticales associatives avec des projections vers le striatum ventromédial et l'hippocampe, tandis que le contrôle « habituel » engage généralement des boucles thalamocorticales sensorimotrices avec projection vers le striatum dorsolatéral et le cervelet (Everitt & Robbins, 2013; Pezzulo, Rigoli et Friston, 2015). Finalement, nous pouvons aussi comprendre la

distinction entre les deux formes d'inférences actives (profonde et superficielle) dans la perspective des stratégies de navigation : Le contrôle « model-based », effectué par les inférences profondes, va correspondre à une stratégie d'apprentissage par renforcement instrumental basée sur une navigation allocentrique (« place-map »), tandis que le contrôle « model-free », effectué par les inférences superficielle, va correspondre à une stratégie d'apprentissage par renforcement instrumental basée sur une navigation égocentrique (« cue-response ») (Anggraini et al., 2018).

Ainsi comprise, l'inférence active permet une sélection de l'action de type « model-based », qui repose sur l'évaluation de valeurs (épistémique et pragmatique) de séquences d'actions concurrentes, mais aussi une sélection d'action de type « model-free », qui repose sur les valeurs de séquences d'action ayant été « mises en cache » (c'est-à-dire, des valeurs précédemment calculées qui sont stockées dans la mémoire procédurale) pour contourner partiellement ou complètement la nécessité de planification (Maisto et al., 2019). Par conséquent, l'inférence active permet d'intégrer le contrôle automatique ou « habituel/pavlovien » et le contrôle délibéré ou « dirigé vers des buts » dans un modèle hiérarchique où les niveaux plus élevés contextualisent les niveaux inférieurs (Friston et al., 2016). Par exemple, les niveaux supérieurs qui encodent des « narratifs sémantiques » sélectionnent des actions « dirigé vers des buts » (par exemple, choisir quoi manger au restaurant), ce qui génère une série de buts et de sous-buts (par exemple, consulter le menu et appeler le serveur) qui vont guider l'action pendant que les niveaux inférieurs qui encodent les « dynamiques sensorimotrice » vont sélectionner des actions « automatiques/habituelles/pavloviennes » sans exiger des buts/sous-buts. Dans le cadre de l'inférence active, les inférences délibérées sont aussi appelée « belief-based » puisqu'elles sélectionnent des actions basées sur une évaluation prospective qui implique des croyances à propos des états futurs et leurs conséquences associées, tandis que les inférences automatiques sont aussi appelées « belief-free » puisqu'elles sélectionnent automatiquement des actions basées sur une comparaison rétrospective des valeurs « mises en cache » et/ou sur des priors adaptatifs spécifiés pas l'évolution, cela en tenant uniquement compte des observations actuelles.

Dans le formalisme de l'inférence active profonde, la planification et l'exécution de l'action impliquent des modèles génératifs avec des propriétés mathématiques différentes de celles que l'on retrouve dans l'inférence perceptuelle et l'inférence active superficielle. La planification de l'action (c'est-à-dire l'évaluation et la sélection de séquences d'actions) implique des états discrets actualisés par un algorithme de propagation de croyances tandis que la perception (c'est-à-dire l'estimation d'états) et le mouvement

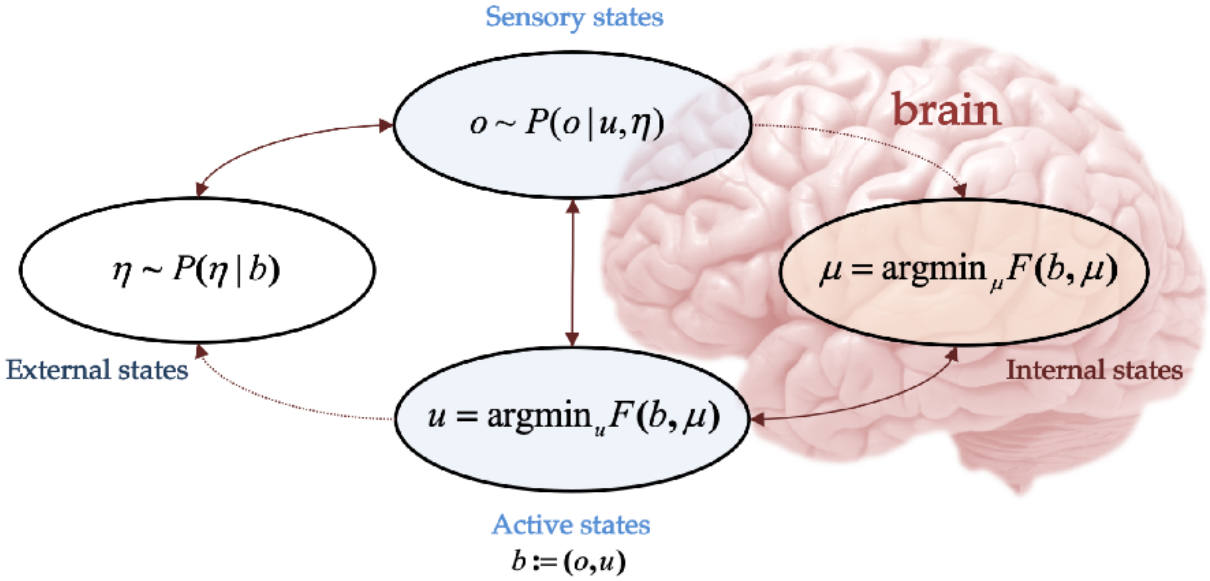
impliquent des états continus actualisés par un algorithme de codage prédictif généralisé. En effet, la perception et le mouvement concret vont dépendre de variables continues (par exemple, le traitement auditif) tandis que la planification d'action abstraite dépend de variables discrètes (par exemple, le traitement langagier) (Friston, Parr et de Vries, 2017). Ceci découle du fait que les modèles génératifs incarnés du cerveau sont organisés hiérarchiquement en fonction des échelles temporelles sur lesquelles leurs représentations évoluent, allant des flux continus et transitoires des « dynamiques sensorimotrices » aux représentations discrètes et stables des « narratifs sémantiques » (Friston & Buzsaki, 2016). Aux niveaux inférieurs, les modèles génératifs continus anticipent l'évolution dynamique des trajectoires sensorimotrices (puisqu'elles sont temporellement retardées en raison des contraintes de conduction axonale), ce qui permet de générer des états cachés continus pouvant se synchroniser avec les événements externes (Perrinet et al., 2014). En termes d'implémentation dans des microcircuits corticaux canoniques, les modèles génératifs continus sont généralement associés aux régions granulaires (sensorielles) et agranulaires (motrices) (Bastos et al., 2012; Shipp et al., 2016). À des niveaux plus élevés, les modèles génératifs discrets génèrent les séquences d'états cachés futurs et leurs conséquences en fonction d'une séquence d'actions (Friston, Parr et de Vries, 2017). En termes d'implémentation dans des microcircuits corticaux canoniques, les modèles génératifs discrets sont typiquement associés aux régions dysgranulaires (associatives) (Friston, Parr et de Vries, 2017). Ici, les inférences discrètes et continues interagissent constamment dans l'inférence active profonde, traduisant des décisions discrètes en mouvements continus et des sensations continues en sémantique discrète à travers une interface (le thalamus moteur pour le système moteur en général et le colliculus supérieur pour le système oculomoteur) (Friston, Parr et de Vries, 2017).

Du point de vue philosophique, plusieurs interprétations contradictoires peuvent être données au formalisme présenté ci-haut. En effet, le FEP peut autant être compris dans la perspective internaliste du cognitivisme classique, qui repose sur des représentations découplées de l'environnement, que dans une perspective externaliste de la cognition incarnée, qui repose sur des interactions couplées à l'environnement. Par exemple, le FEP décrit la frontière des systèmes biologiques comme des couvertures de Markov[36] divisées en état externe, état sensoriel, état interne et état actif (figure 5). Les dépendances entre ces états induisent une relation causale qui préserve la frontière de la couverture de Markov des

[36] Il est important de distinguer entre les couvertures de Markov de Pearl (1998) et celles de Friston (2010), ces dernières prenant la forme d'un graphe cyclique dirigé (ou dynamic causal modelling) qui capture les indépendances conditionnelles entre les noeuds d'un réseau Bayésien.

systèmes biologiques en générant des boucles action-perception qui minimisent l'énergie libre. Pour les internalistes, le FEP implique généralement l'isolement inférentiel de la couverture de Markov du cerveau, limitant la cognition au système nerveux (Hohwy, 2017). Selon cette perspective, les modèles génératifs représentationnels effectuent des inférences sur des états cachés impliquant du contenu qui concernent des choses dans le monde (Gładziejewski et Miłkowski, 2017; Kiefer & Hohwy, 2018). Pour les externalistes, la FEP implique l'ouverture inférentielle de la couverture de Markov du cerveau, étendant la cognition au corps et à l'environnement (Bruineberg, Kiervertein et Rietveld, 2018; Kirchhoff & Robertson, 2016). Selon cette perspective, les modèles génératifs non-représentationnels effectuent des inférences sur des états cachés couplés dynamiquement qui régulent l'engagement dirigé vers le monde (Kirchhoff & Robertson, 2018; Ramstead, Kirchhoff & Friston, 2020). Cependant, il existe également un compromis qui soutient que le principe de l'énergie libre fournit une perspective intégrative qui pourrait potentiellement dissoudre les tensions entre l'internalisme et l'externalisme (Allen & Friston, 2016; Ramstead et al, 2019; Constant, Clark et Friston, 2021). Plutôt que d'assimiler le formalisme du FEP aux interprétations internalistes ou externalistes, nous soutiendrons dans les prochaines sections qu'il serait plus avantageux d'envisager comment le FEP permet de dissoudre les tensions théoriques dans un pluralisme explicatif intégratif qui implique une interaction bidirectionnelle entre aspects fonctionnalistes (parties au tout) et structuralistes (tout aux parties) de l'architecture. En effet, nous pouvons comprendre le processus d'auto-organisation globale par minimisation de l'énergie libre comme une explication structuraliste (explication formelle, non-causale; voir Huneman, 2018) de l'architecture cognitive. Toutefois, la minimisation de l'énergie libre implémente aussi diverses composantes fonctionnelles locales que nous pouvons comprendre dans la perspective d'une explication fonctionnaliste (explication causale-mécaniste; voir Craver, 2007) de l'architecture cognitive (notamment en termes d'implémentation dans des microcircuits canoniques locaux compris comme des schémas mécanistes). C'est dans cette perspective que nous abordons maintenant l'esprit hiérarchiquement mécaniste (HMM), une proposition d'architecture cognitive sous le principe d'énergie libre (Badcock et al., 2019).

Figure 5 : Réprésentation de l'inférence active sous une couverture de Markov



Le schéma illustre la séparation entre des états internes et externes par une couverture de Markov qui comprend des états sensoriels et actifs. Dans le cerveau, la partition implique une distinction entre action et perception qui minimise l'énergie libre dans des boucles synergiques. La minimisation de l'énergie libre entraîne l'auto-organisation des états internes par la perception (inférence perceptuelle) tandis que l'action (inférence active) entraîne le couplage du cerveau aux états externes (Tirée Da Costa et al., 2020).

2.2 L'architecture cognitive sous l'esprit hiérarchiquement mécaniste

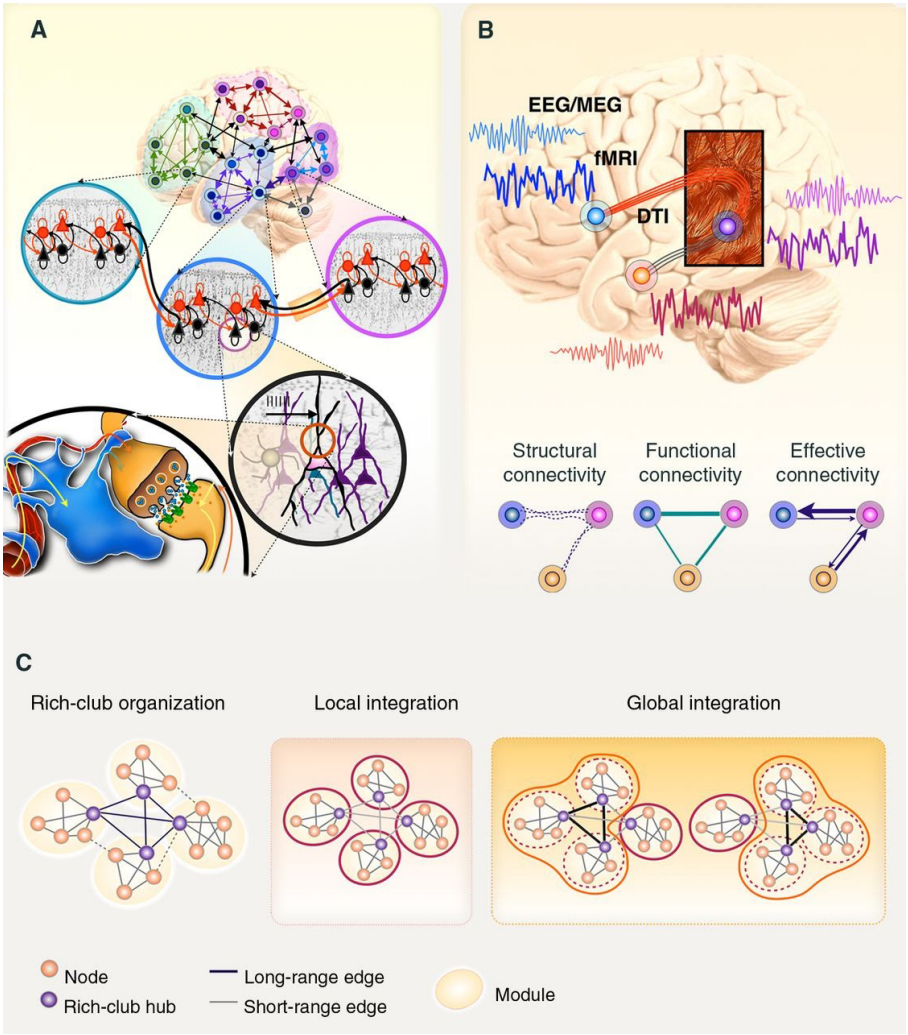
Dans cette section, nous présentons la HMM, une proposition d'architecture cognitive sous le FEP (Badcock, 2012, Badcock et al., 2019), dans la perspective d'un pluralisme explicatif intégratif qui distingue entre aspects fonctionnalistes (sélectionnistes) et structuralistes (développementalistes). En résumé, la HMM soutient que le cerveau est un système adaptatif complexe qui minimise activement l'entropie de ses états internes par des cycles action-perception générés par des interactions dynamiques bidirectionnelles entre des mécanismes neurocognitifs hiérarchiquement organisés, allant des mécanismes « domaine-spécifiques » hautement ségrégués des niveaux inférieurs aux mécanismes « domaine-généraux » hautement intégrés des niveaux supérieurs (Badcock et al., 2019). Dans le cadre de notre proposition qui repose sur un pluralisme explicatif intégratif, nous proposons une distinction entre deux formes d'explications dans la HMM: (1) Une explication « structuraliste » (formelle, non-causale) associée aux dynamiques neuronales hiérarchiques qui minimisent l'énergie libre et (2) une explication « fonctionnaliste » (causale-mécaniste) associée aux mécanismes neurocognitifs hiérarchiquement organisés qui sous-tendent les dynamiques neuronales. D'un point de vue structuraliste, la HMM soutient

que les processus cognitifs émergent des dynamiques neuronales hiérarchiques qui minimisent l'énergie libre variationnelle par transmission bidirectionnelle de messages entre mécanismes neurocognitifs. D'un point de vue fonctionnaliste, les mécanismes neurocognitifs sont définis comme des sous-systèmes neuronaux sur plusieurs échelles spatiales qui sont caractérisés par: (1) une ségrégation fonctionnelle locale (dans des modules des réseaux) par des connections denses de courtes portées et (2) une intégration fonctionnelle globale (dans des hubs connecteurs des réseaux) par des connections éparses de longues portées (Park & Friston, 2013). Dans cette perspective, « [...] cognition can be described as the global integration of local (segregated) neuronal operations that underlies hierarchical message passing among cortical areas, and which is facilitated by hierarchical modular network architectures » (Park & Friston, 2013, p. 579). Selon la HMM, une propriété essentielle des sous-systèmes neuronaux est leur quasi-décomposabilité hiérarchique (voir Simon, 1991): Les sous-systèmes neuronaux qui sont localement ségrégués par des connexions de courtes portées demeurent globalement intégrés par des connexions de longue portée, ce qui facilite la transmission de message bidirectionnelle (Badcock et al., 2019).

Cette conception de l'architecture cognitive du cerveau, qui dépend de l'expression mutuelle de la ségrégation fonctionnelle locale (modulaire) et de l'intégration fonctionnelle globale (intermodulaire), reçoit un certain degré de confirmation empirique dans les neurosciences de réseau. Les neurosciences des réseaux modélisent les réseaux neuronaux du cerveau en utilisant la théorie des graphes pour représenter des nœuds (c'est-à-dire les éléments neuronaux) et des arêtes (c'est-à-dire les connexions entre les éléments neuronaux) (Sporns, 2010). Les analyses de la connectivité (structurelle, fonctionnelle et effective) des réseaux cérébraux sur différentes échelles spatiales vont révéler une organisation hiérarchiquement modulaire interconnectée par des hubs « rich-club » (figure 6) (Park & Friston, 2013). La notion de hiérarchie qui est pertinente ici est celle d'une hiérarchique auto-similaire qui repose sur l'encapsulation répétée d'éléments neuronaux plus petits dans des plus grands, c'est-à-dire qu'un noeuds (réseaux, modules ou sous-modules) comprend un réseau de noeuds interactifs plus petits à un niveau inférieur (grands réseaux neuronaux, régions neuronales, colonnes corticales, neurones) (Meunier et al, 2010; Park & Friston, 2013). Dans le formalisme des neurosciences des réseaux, les modules des réseaux désignent des nœuds localement interconnectés par des arêtes de courte portée qui réalisent une fonction spécialisée (ségrégation locale). Les hubs connecteurs désignent des nœuds globalement interconnectés par des arêtes de longue portée qui facilitent la communication intermodulaire (intégration globale). Quant aux hubs « rich-club », ils désignent des hubs connecteurs qui sont fortement interconnectés à d'autre hubs connecteurs par des connexions de longue portée. Selon Park et Friston (2013), les hubs

« rich-club » supportent la diversité fonctionnelle en fournissant des connexions de longue portée qui permettent une intégration intermodulaire sensible au contexte, laquelle serait supportée par la modulation du gain synaptique et les interactions non-linéaires synchronisées. En accord avec cette proposition, les analyses des réseaux ont révélées que l'activité des hubs connecteurs (qui sont situées dans des régions fonctionnellement diversifiées) augmente lors de tâches impliquant plusieurs fonctions cognitives tandis que celle des modules n'augmentent pas, ce qui indique que les hubs connecteurs permettent la coordination de la connectivité effective intermodulaire en maintenant l'indépendance fonctionnelle des modules (qui sont situés dans des régions fonctionnellement spécialisées) (Bertolero, Yeo et D'Esposito, 2015). Toutefois, déterminer avec précision dans quelle mesure une région neuronale est fonctionnellement ségréguée et intégrée demeure ultimement une question empirique qui dépend des avancées techniques en analyse de la connectivité.

Figure 6 : Organisation hiérarchiquement modulaire des réseaux neuronaux



A : L'organisation multi-échelle des réseaux du cerveau : des neurones, aux colonnes corticales, jusqu'aux régions macroscopiques. Les réseaux sont composés de nœuds et d'arêtes. Les nœuds sont des unités interactives du réseau qui sont eux-mêmes composés de nœuds interactifs plus petits aux niveaux inférieurs. B : Les arêtes des réseaux sont définies par trois types de connectivités : connectivité structurelle, connectivité fonctionnelle et connectivité effective. La connectivité structurelle, qui caractérise les connexions anatomiques, est estimée par IRM de diffusion. La connectivité fonctionnelle, qui caractérise la dépendance statistique entre des activations neuronales, est estimée par IRMf-BOLD et EEG/MEG. La connectivité effective, qui caractérise les interactions causales relatives à une tâche et à des flux d'information à partir d'un modèle de l'interaction neuronale (comme le « dynamic causal modeling »), est aussi estimée par IRMf-BOLD et EEG/MEG. C : Les analyses de connectivité sur différentes échelles spatiales vont révéler une organisation hiérarchiquement modulaire interconnectée par des hubs « rich-club ». Dans les neurosciences des réseaux, les modules sont des descriptions topologiques qui consiste en des ensembles de nœuds densément interconnectés par des connexions de courtes portées (intégration locale), mais qui restent tout de même faiblement connectés aux nœuds du reste du réseau. Les hubs connecteurs sont des nœuds densément

> interconnectés par des connexions de longues portées qui permettent l'interconnexion modulaire (intégration globale). Les hubs « rich-club » sont quant à eux des hubs qui sont densément interconnectés à d'autres hubs par des connexions de longues portées (Tirée de Park & Friston, 2013).

De retour dans la perspective du FEP, nous pouvons maintenant aborder la HMM comme une explication neurobiologiquement plausible pour la ségrégation et l'intégration fonctionnelle dans les réseaux hiérarchiquement modulaires du cerveau. Comme nous l'avons abordé dans la section précédente, les systèmes biologiques sont compris sous cette perspective comme des modèles génératifs incarnés qui sont optimisés pour refléter de la structure statistique de l'environnement sur plusieurs échelles temporelles. En ce sens, les cerveaux ne font pas seulement qu'encoder des modèles génératifs, ils sont une réflexion physique des régularité statistiques causales de l'environnement ayant été optimisé par l'auto-organisation et la sélection naturelle (de modèles bayésiens incarnés). Cette perspective découle du théorème du bon régulateur en cybernétique qui soutient que tout système capable de réguler son environnement doit instancier un bon modèle de cet environnement (Conant & Ashby, 1970; Friston & Buzsaki, 2016). Par conséquent, on peut s'attendre à ce que la structure anatomique du cerveau reflète la structure causale de l'environnement sur plusieurs échelles temporelles. Plus particulièrement, nous pouvons distinguer entre les échelles temporelles du fonctionnalisme sélectionniste, qui permettent d'expliquer la ségrégation locale modulaire aux niveaux inférieurs, et du structuralisme développementaliste, qui permettent d'expliquer l'intégration globale intermodulaire aux niveaux supérieurs. Dans les sections suivantes, nous soutiendrons que les pressions sélectives locales vont façonner des modules « domaine-spécifiques » aux niveaux inférieurs et les pressions sélectives globales vont façonner des hubs connecteurs « rich-club » « domaine-généraux » aux niveaux supérieurs (voir Lundie, 2019).

Du point de vue de la ségrégation fonctionnelle[37], la HMM soutient qu'elle émerge des régularités statistiques conservées sur des échelles temporelles évolutionnaires. Formellement, la ségrégation fonctionnelle émerge d'un découpage statistique approximatif de « la nature à ses joints » à travers diverses « approximations de champ moléculaires »[38] (mean-field approximation) qui factorisent les

[37] Dans les neurosciences des réseaux, la ségrégation fonctionnelle réfère à la spécialisation d'une région du cerveau pour des fonctions cognitives ou sensorimotrices particulières, ce qui implique la ségrégation anatomique de la fonction dans une région neuronale (Friston & Price, 2011).

[38] Les approximations du champ moléculaire permettent une approximation d'une distribution conjointe sur une ou plusieurs variables aléatoires avec un produit de leurs distributions marginales (Friston & Buzsaki, 2016).

représentations des causes cachées des sensations (Friston & Buzsaki, 2016). En somme: « [...] evolutionary (Bayesian belief) updates have shaped the brain into an efficient (minimum free energy) mean field approximation that we know and study as functional segregation » (Friston & Buzsaki, 2016, p. 502). Par exemple, l'indépendance statistique (indépendance conditionnelle) entre les causes cachées associées à l'identité des objets par rapport à celles associées à la localisation des objets visuels suggère une séparation anatomique (absence d'arêtes entre des noeuds) entre les représentations du « quoi », associée à la voie ventrale, et du « où/comment », associée à la voie dorsale (Ungerleider & Mishkin, 1982). Ceci suggère que les causes environnementales indépendantes sont encodées dans des structures neuronales fonctionnellement ségréguées (nous pouvons connaître l'identité d'un objet indépendamment de sa localisation dans l'espace et vice versa) (Friston & Buzsaki, 2016). Fait intéressant, il existerait également une séparation anatomique entre les représentations du « quoi » et du « quand » dans la voie ventrale (Friston & Buzsaki, 2016). En effet, le cervelet et l'hippocampe seraient des structures qui encodent des successions temporelles tandis que les cortex sensorimoteurs et associatifs encodent les représentations (ou contenus) des séquences. Dans cette perspective, l'hippocampe est un hub connecteur qui entretient de nombreuses connexions convergentes et divergentes avec les représentations de contenu du « quoi » d'un néocortex plus modulaire, ce qui place l'hippocampe et le cortex paralimbique au centre d'une hiérarchie centrifuge du cerveau (Friston & Buzsaki, 2016). Ainsi, l'architecture cognitive du cerveau serait une transcription statistique de la structure hiérarchique du monde, ce qui entraîne une séparation hiérarchique des échelles temporelles auxquelles les représentations évoluent, des fluctuations rapides des « dynamiques sensorimotrices » (représentations continues du « comment » et du « où ») aux changements contextuels lents des « narratifs sémantiques » (représentations discrètes du « quoi » et du « quand »).

Du point de vue de l'intégration fonctionnelle[39], la HMM soutient qu'elle émerge de la transmission bidirectionnelle de message entre régions neuronales, un processus facilité par l'organisation hiérarchiquement modulaire du cerveau. Formellement, la transmission de message hiérarchique est généralement formulée en termes de propagation de croyances pour les niveaux discrets (planification abstraite) et en termes de codage prédictif généralisé pour les niveaux continus (exécution concrète) (voir Parr, de Vries et Friston, 2017). Lorsqu'on l'exprime en termes d'erreurs de prédiction, le filtrage bayésien correspond au codage prédictif qui, comme nous l'avons précédemment mentionné dans la section

[39] Dans les neurosciences des réseaux, l'intégration fonctionnelle réfère au couplage et interactions (transmission de messages) entre régions neuronales (Friston & Price, 2011).

précédente, est subsumé par le formalisme du FEP. Dans cette perspective, la minimisation d'erreurs de prédictions implique un mécanisme d'ajustement dynamique (en temps réel) de la connectivité (poids des arêtes) dans les réseaux neuronaux en modifiant l'efficacité synaptique par l'entremise des connexions qui acheminent des prédictions aux niveaux inférieurs et d'autres qui acheminent des erreurs de prédictions aux niveaux supérieurs (Park & Friston, 2013). Les états intrinsèques des noeuds ou modules et les poids des arêtes vont être récursivement actualisés pour améliorer les prédictions sur chacun des niveaux de la hiérarchie, tandis que le poids des arêtes dirigées (dans un graphe dirigé qui reflète la direction de l'information) va correspondre à la connectivité effective d'un réseau (la relation causale dirigée entre les noeuds ou modules) lors d'une tâche spécifique. La cognition peut alors être comprise comme l'intégration globale d'opérations neuronales localement ségréguées par transmission de message hiérarchique de minimisation d'erreurs entre différentes régions corticales via l'architecture hiérarchiquement modulaire des réseaux du cerveau (Park & Friston, 2013). Du point de vue des dynamiques neuronales, la transmission de message dans les réseaux hiérarchiquement modulaires favorise la « criticalité autoorganisée » (Bak et al., 1987), une propriété dynamique que possèdent les systèmes adaptatifs complexes qui vont fluctuer entre des états hautement ordonnés (ségrégués) et des états hautement désordonnés (intégrés). En effet, l'équilibre adéquat entre la ségrégation et l'intégration dans l'organisation modulaire hiérarchique semble élargir la gamme des paramètres de la criticalité autoorganisée (Hilgetag & Hutt, 2014). Le maintien d'une coexistence entre dynamiques sous-critiques (hautement ordonnées/ségréguées) et supercritiques (hautement désordonnées/intégrées) à différents niveaux hiérarchiques est probablement une caractéristique retenue par la sélection naturelle car elle semble favoriser un traitement optimal de l'information (Hesse & Gross, 2014). Ce faisant, l'architecture hiérarchiquement modulaire du cerveau semble favoriser le maintien d'un état critique « à la frontière du chaos » qui permet de basculer entre différentes dynamiques neuronales en fonction des perturbations (Hesse & Gross, 2014). Puisque cet état de criticalité autoorganisée situé entre l'ordre et le chaos serait riche sur le plan informationnel, il serait une signature des dynamiques neuronales « sensibles au contexte » qui sous-tendent la conscience[40] (Taggliazucchi, 2017).

[40] Dans le modèle de l'espace de travail neuronal global de Baars, Changeux et Daheane (2005), des processeurs locaux compétitionnent en parallèle dans la périphérie neuronale, ce qui peut entrainer une transition critique nommé « ignition » lorsque les interactions non-linéaires sont suffisantes pour permettre l'accès à l'ensemble de neurones distribués qui constituent l'espace de travail.

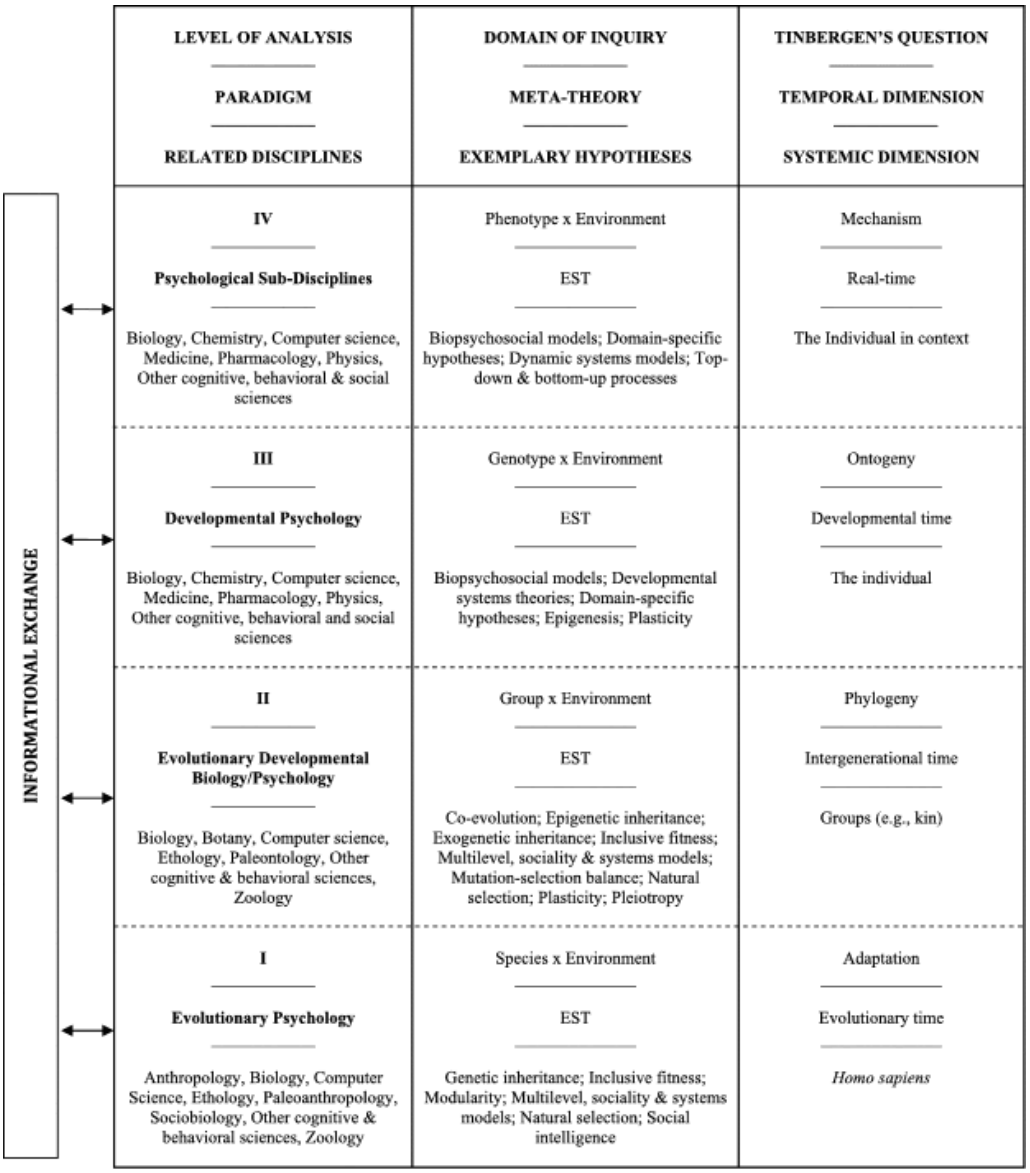
En plus de fournir une explication de la ségrégation et de l'intégration fonctionnelle dans les réseaux neuronaux du cerveau, la HMM fournit aussi une explication de l'évolution de l'architecture cognitive sous la EST. Fondamentalement, la EST repose sur l'interaction réciproque entre la sélection et l'auto-organisation pour expliquer la fonction et la forme de l'architecture cognitive du cerveau. La sélection fait référence aux principes interactifs du changement - variation, sélection et rétention - qui opèrent sur des systèmes dynamiquement couplés (Caporael, 2001). Dans cette perspective, la sélection est un processus universel qui s'étend à tous les niveaux d'organisation (physique, chimique, biologique, psychologique, socioculturel) pour incorporer tous les systèmes dynamiquement couplés. On qualifie cette sélection de « naturelle » lorsqu'elle porte sur les systèmes biologiques, mais on peut aussi parler de sélection « générale » lorsqu'elle porte sur d'autres types de systèmes (physiques, chimiques, culturels, etc.). L'auto-organisation fait référence à l'émergence spontanée de patrons cohérents d'ordre supérieur à partir d'interactions récursives entre des composants plus simples d'un système dynamique complexe (Lewis, 2000). Dans cette perspective, l'auto-organisation est caractérisée par quatre propriétés importantes (Badcock, 2012, p.14-15):

> First, following some sort of (internal or external) environmental perturbation or trigger, microscopic coordinations emerge between different components that lead to new macroscopic patterns, which are recruited for unique functions that entrain and reinforce certain lower-order patterns over time (a process of circular causality between different levels of the system). Second, they become progressively complex and ordered, producing increasingly sophisticated arrangements of coordinated processes to subserve increasingly functional forms. Third, global reorganizations toward increasing complexity occur at phase transitions: points of turbulent instability where old patterns break down to be replaced by new ones. Fourth, they are both stable and sensitive to change: emergent change is stabilized through negative feedback loops and the coordination of functional patterns, while an interconnectedness with other systems and propensity for positive feedback favors sensitivity to environmental conditions, particularly during phase transitions. Thus, self-organizing processes tend toward nonlinearity.

En résumé, la EST soutient que la sélection et l'auto-organisation sont des processus fondamentaux qui se renforcent mutuellement puisque des patrons globaux autoorganisés de composantes locales en interactions sont sélectionnés parmis des alternatives pour permettre l'émergence de différents niveaux d'organisation (Kauffman, 1993, Badcock, 2012, Badcock et al., 2019). Ce faisant, il devient possible de construire une théorie des systèmes évolutionnaires au carrefour du fonctionnalisme, qui fournit une théorie de la fonction locale, et du structuralisme, qui fournit une théorie de la fonction globale.

Plus particulièrement, la EST explique la fonction et la forme de l'architecture hiérarchiquement modulaire du cerveau comme résultant d'interactions dynamiques qui minimisent l'énergie libre sur les quatre niveaux d'analyse biologique (associées aux quatre questions de Tinbergen; 1963) (figure 7): (1) l'adaptation (fonction) qui concerne les explications évolutionnaires des caractéristiques « typiques à l'espèce », ce qu'on associe à la psychologie évolutionniste et à la synthèse moderne néo-darwinnienne. Ce niveau implique l'optimisation de l'énergie libre moyenne au cours de l'évolution par l'influence des pressions sélectives sur les modèles génératifs (en y insérant des priors adaptatifs); (2) la phylogénie (évolution) qui concerne les explications intergénérationnelles des similitudes et différences entre groupes, ce qu'on associe à la psychologie/biologie évolutionnaire-développementale et la synthèse évolutionnaire étendue. Ce niveau implique l'optimisation de l'énergie libre moyenne au cours des générations par transmission épigénétique/exogénétique des modèles génératifs (ici encore en termes de priors adaptatifs); (3) l'ontogénie (développement) qui concerne les explications développementales des caractéristiques des individus, ce qu'on associe à la psychologie développementale. Ce niveau implique l'optimisation des modèles génératifs (priors et fonctions de vraisemblance) par l'émondage dépendant de l'activité et le maintien des connections neuronales transmisent épigénétiquement; (4) le mécanisme (causalité) qui concerne les explications en temps réels des phénomènes neurocognitifs, ce qu'on associe aux sous-disciplines de la psychologie (psychologie cognitive, psychologie de l'apprentissage, neuropsychologie, etc.). Ce niveau implique l'optimisation de l'action et de la perception par la minimisation de l'énergie libre des modèles génératifs et l'optimisation du gain synaptique qui encode la précision dans l'attention et l'apprentissage (Badcock, 2012; Badcock et al., 2019). La HMM repose sur des « priors adaptatifs » génétiquement et épigénétiquement hérités qui ont été façonnés par les pressions sélectives pour guider les cycles action-perception vers des états adaptatifs (i.e., ceux qui ont peu de « surprise »). En ce sens, la sélection naturelle permet de minimiser la « surprise » en attribuant une valeur innée aux états sensoriels par transmission génétique et épigénétique de modèles génératifs (priors adaptatifs). Comme nous l'avons abordé au premier chapitre, nous proposons de comprendre les niveaux d'analyse biologique dans la perspective d'un pluralisme explicatif intégratif qui distingue entre explications fonctionnalistes (parties au tout) et explications structuralistes (tout aux parties). Ce faisant, nous serons en mesure de comprendre comment la sélection naturelle et l'auto-organisation interagissent pour façonner l'architecture cognitive.

Figure 7 : La théorie des systèmes évolutionnaires



LEVEL OF ANALYSIS PARADIGM RELATED DISCIPLINES	DOMAIN OF INQUIRY META-THEORY EXEMPLARY HYPOTHESES	TINBERGEN'S QUESTION TEMPORAL DIMENSION SYSTEMIC DIMENSION
IV **Psychological Sub-Disciplines** Biology, Chemistry, Computer science, Medicine, Pharmacology, Physics, Other cognitive, behavioral & social sciences	Phenotype x Environment EST Biopsychosocial models; Domain-specific hypotheses; Dynamic systems models; Top-down & bottom-up processes	Mechanism Real-time The Individual in context
III **Developmental Psychology** Biology, Chemistry, Computer science, Medicine, Pharmacology, Physics, Other cognitive, behavioral and social sciences	Genotype x Environment EST Biopsychosocial models; Developmental systems theories; Domain-specific hypotheses; Epigenesis; Plasticity	Ontogeny Developmental time The individual
II **Evolutionary Developmental Biology/Psychology** Biology, Botany, Computer science, Ethology, Paleontology, Other cognitive & behavioral sciences, Zoology	Group x Environment EST Co-evolution; Epigenetic inheritance; Exogenetic inheritance; Inclusive fitness; Multilevel, sociality & systems models; Mutation-selection balance; Natural selection; Plasticity; Pleiotropy	Phylogeny Intergenerational time Groups (e.g., kin)
I **Evolutionary Psychology** Anthropology, Biology, Computer Science, Ethology, Paleoanthropology, Sociobiology, Other cognitive & behavioral sciences, Zoology	Species x Environment EST Genetic inheritance; Inclusive fitness; Modularity; Multilevel, sociality & systems models; Natural selection; Social intelligence	Adaptation Evolutionary time *Homo sapiens*

Note: Adapted from Badcock (2012)

L'architecture cognitive du cerveau émerge de l'influence complémentaire de la sélection (générale et naturelle) et de l'auto-organisation qui agissent sur les quatre niveaux d'analyse biologique dynamiquement couplés : adaptation, phylogénie, ontogenèse, et mécanisme. Les paradigmes en psychologie illustrent ce processus en se concentrant différemment sur quatre niveaux spécifiques et interdépendants d'analyse : explications fonctionnelles pour les caractéristiques adaptatives « typiques de l'espèce » (c'est-à-dire la psychologie évolutionniste et la synthèse moderne); explications intergénérationnelles pour les similitudes et différences entre les groupes (c'est-à-dire la psychologie évo-dévo et la synthèse évolutionnaire étendue); explications pour le développement individuel (c'est-à-dire la psychologie du développement); et explications mécanistes pour les phénomènes neurocognitifs en temps réel (c'est-à-dire les sous-disciplines en psychologie) (tirée de Badcock et al., 2019).

2.2.1 L'architecture cognitive computationnelle et dynamique sous l'inférence active profonde

Dans le premier chapitre, nous avons vu que l'hypothèse de modularité massive (MMH) décrit le cerveau comme un système computationnel de traitement de l'information. La cognition reposerait alors sur la manipulation séquentielle des représentations découplées de l'environnement (Carruthers, 2006). Par ailleurs, nous avons aussi vu que l'hypothèse du redéploiement massif (MRH) décrit le cerveau comme un système de contrôle dynamique. La cognition dépendrait dans ce cas de la compétition parallèle d'affordances qui émerge d'interactions couplées avec l'environnement (Anderson, 2016). Dans cette section, nous discutons de la manière dont l'inférence active profonde permet d'envisager la complémentarité entre ces deux hypothèses en distinguant entre les processus « délibérés » ou « dirigés vers des buts » aux niveaux supérieurs et les processus « automatiques » ou « habituels/pavloviens » aux niveaux inférieurs.

Selon notre proposition, les processus « délibérés » ou « dirigés vers des buts » vont impliquer une manipulation de représentations découplées de l'environnement au cours de la planification de l'action tandis que les processus « automatiques » ou « habituels/pavloviens » vont impliquer des interactions sensorimotrices couplées avec l'environnement qui permettent l'exécution experte ou réflexe de l'action. Dans la perspective d'un pluralisme explicatif intégratif, nous soutiendrons que la gestion des interactions couplées avec l'environnement repose sur un processus dynamique d'auto-organisation globale d'un modèle dans le système cerveau-corps-environnement (explication structuraliste) tandis que la manipulation des représentations découplées de l'environnement implique additionnellement un processus computationnel de sélection locale de modèles dans le système nerveux (explication fonctionnaliste). Suivant l'inférence active profonde, c'est le degré d'incertitude contextuelle détermine la profondeur hiérarchique du processus d'inférence, des inférences « automatiques » ou « habituelles/pavloviennes » aux inférences « délibérées » ou « dirigées vers des buts » (Friston et al., 2016; Pezzulo et al., 2018). Plus précisément, nous soutiendrons que la sélection d'action est un processus d'inférence hiérarchique qui peut impliquer des inférences profondes « avec croyances » (ou « goal-driven »), qui dépendent de l'évaluation de la valeur épistémique et pragmatique pour sélectionner la séquence d'actions optimale en présence d'incertitude contextuelle, et des inférences superficielles « sans croyances » (ou « stimulus-driven »), qui dépendent de valeurs précédemment mises en cache ou des priors adaptatifs pour sélectionner la séquence d'action optimale en l'absence relative d'incertitude contextuelle (Friston et al., 2016; Maisto et al., 2019).

Dans l'inférence active profonde, la planification de l'action délibérée implique des états discrets puisqu'il faut sélectionner entre différentes séquences d'actions contrefactuelles en fonction d'une évaluation de leur énergie libre attendue. En d'autres termes, les inférences actives profondes reposent surtout sur des modèles génératifs discrets qui vont simuler la transition des états sensoriels futurs et de leurs conséquences associées sous des séquences d'actions alternatives. Les modèles génératifs discrets concernent les conséquences de l'action qui sont causées par des états cachés décrivant le futur et qui par définition ne peuvent donc pas être directement observés. Ainsi, la planification de l'action implique l'évaluation des séquences d'actions en fonction de croyances concernant les conséquences attendues ou préférées (valeur pragmatique ou motivationnelle), lesquelles confèrent à l'action planifiée son caractère « dirigé vers des buts ». C'est pourquoi nous avons mentionné que la planification de l'action implique des inférences « avec croyances », c'est-à-dire qu'elles impliquent l'évaluation de l'énergie libre attendue sous différentes séquences d'actions. Plus précisément, elles impliquent la minimisation du risque (valeur pragmatique), c'est-à-dire la capacité inférée qu'aurait une séquence d'action à réaliser une conséquence attendue ou préférée, et la minimisation de l'ambiguïté attendue (valeur épistémique), c'est-à-dire la capacité inférée qu'aurait une séquence d'action de produire des conséquences informatives. D'un point de vue plus formel, les inférences profondes « avec croyances » des modèles génératifs discrets vont surtout reposer sur des processus de décision de Markov qui spécifient la trajectoire des états futurs (transition d'états) sous des séquences d'actions alternatives et leurs conséquences associées (Friston, Parr et de Vries, 2017).

Dans l'inférence active superficielle, l'exécution experte ou réflexe de l'action implique des états continus puisqu'il faut optimiser les trajectoires des séquences d'actions (dont les valeurs ont été précédemment calculées puis mises en cache) en fonction des observations sensorielles pour minimiser l'énergie libre variationnelle. En d'autres termes, les inférences actives superficielles reposent surtout sur des modèles génératifs continus pour réaliser les séquences d'actions puisque les sensations et mouvements évoluent dynamiquement dans un monde physique. Les modèles génératifs continus concernent les conséquences de l'action qui sont causées par des états cachés actuels qui peuvent être observés directement. Ainsi, l'exécution experte ou réflexe de l'action implique des valeurs de séquences d'actions « mises en cache » (ou stockées en mémoire procédurale) ou des valeurs prescrites par des priors adaptatifs. Elles ne vont pas nécessiter de croyances concernant les états et conséquences futures, ce qui permet de conférer un caractère « habituel » ou réflexe à l'action. C'est pourquoi nous avons mentionné que l'action habituelle/pavlovienne implique des inférences « sans croyances », c'est-à-dire qu'elles contournent

l'évaluation de l'énergie libre attendue puisqu'il suffit d'inférer les états (cachés) actuels et d'exécuter l'action appropriée au contexte. Plus précisément, elles impliquent la minimisation de l'énergie libre variationnelle par l'exécution automatique de séquences d'actions en fonction d'une association stimulus-réponse. D'un point de vue formel, les inférences superficielles « sans croyances » des modèles génératifs continus vont surtout reposer sur le codage prédictif généralisé qui spécifie l'évolution dynamique des sensations et mouvements (avec quelques fluctuations aléatoires) (Friston, Parr et de Vries, 2017).

Selon notre proposition, les modèles génératifs discrets associés aux inférences « dirigées vers des buts » des niveaux supérieurs peuvent être interprétés du point de vue computationnel et représentationnel de la MMH. Nous soutenons que l'inférence active profonde « avec croyances » peut être comprise comme un processus de manipulation séquentielle de représentations découplées de l'environnement. Comme nous l'avons abordé dans le premier chapitre, les représentations sont traditionnellement définies par un contenu sémantique qui indique quelles choses dans le monde elles représentent (ce à propos de quoi elles sont). Nous proposons d'envisager cette intentionnalité représentationnelle ou cognitive comme étant une caractéristique essentielle des processus « dirigés vers des buts ». Les processus « dirigées vers des buts » dépendent des représentations des buts futurs, lesquelles possèdent des propriétés fonctionnelles particulières. Dans cette perspective, les représentations sont découplées de l'environnement pour permettre la minimisation de l'énergie libre attendue par l'attribution de valeurs de motivation (pragmatique) et de contrôle (épistémique). Pour y parvenir, les processus « dirigés vers des buts » génèrent des boucles action-perceptions découplées de l'environnement pour évaluer l'énergie libre attendue durant la planification. En ce sens, nous proposons de caractériser la notion de représentation dans la perspective d'une propriété fonctionnelle particulière, c'est-à-dire l'utilisation « offline » des modèles génératifs pour permettre l'apprentissage vicariant par essais et erreur. En effet, les modèles génératifs discrets permettent l'évaluation contrefactuelle des séquences d'actions alternatives en fonction de leurs capacités à réaliser des buts, lesquels sont compris comme les conséquences attendues ou préférées de l'action. Ce faisant, les processus « dirigés vers des buts » peuvent être compris comme impliquant des représentations qui sont « à propos » de quelque chose puisqu'ils impliquent des croyances à propos des états cachés du monde (association états-conséquences). Les processus « dirigés vers des buts » possèdent alors des « conditions de satisfaction » puisqu'on peut les évaluer selon qu'ils permettent de « bonnes » et « mauvaises » séquences d'actions en fonction de l'énergie libre attendue.

D'un point de vue neurobiologique, les processus « dirigés vers des buts » vont engager des boucles thalamocorticales associatives avec des projections vers le striatum ventromédial/dorsomedial et l'hippocampe (Everitt & Robbins, 2013; Pezzulo, Rigoli et Friston, 2015). Ces boucles associent les représentations (sémantiques) discrètes de la voie du « quoi » et les représentations (épisodiques) discrètes de la voie du « quand » pour générer des « narratifs sémantiques » (Buzsaki & Friston, 2016). Les « narratifs sémantiques » sont des séquences hippocampiques générées à l'interne qui permettent la cognition orientée vers le futur (et le passé, puisque c'est le passé qui permet de prédire le futur dans un monde qui repose sur des lois de la nature) (Pezzulo et al., 2017). Dans la perspective de la MMH, cette conception des processus « dirigés vers des buts » est associée aux boucles de répétition mentale consciente de schémas d'actions qui sont diffusées dans « l'espace de travail global », ce qui permet de séquentiellement coopter des modules centraux de croyances et de désirs (Carruthers, 2006). Comme nous l'avons vu dans le premier chapitre, la répétition mentale consciente repose sur des contenus propositionnels qui sont formulés dans un langage de la pensée (LOT), caractérisé par une structure compositionnelle. Dans la variante peu exigeante du LOT proposée par Carruthers (2006), nous pouvons incorporer n'importe quels modèles qui reposent sur une structure compositionnelle, comme ceux qui permettent la planification ou la navigation allocentrique par exemple. Nous pouvons aisément envisager les « narratifs sémantiques » en termes de séquences hippocampiques qui permettent d'intégrer les contenus sémantiques dans des ensembles compositionnellement structurés.

Selon notre proposition, les modèles génératifs continus associés à des inférences « habituelles/pavloviennes » aux niveaux inférieurs peuvent être interprétées du point de vue dynamique non-représentationnel de la MRH. Nous soutenons que l'inférence active superficielle « sans croyances » dépend d'un processus de compétition parallèle d'affordances (Cisek, 2007) qui émergent des interactions couplées avec l'environnement. Comme nous l'avons soutenu dans le premier chapitre, les interactions peuvent être définies par une tendance dirigée vers une « emprise optimale » (optimal grip) sur un « champ d'affordances ». Nous proposons d'envisager cette intentionnalité experte ou motrice comme étant une caractéristique essentielle des processus « habituels ». En effet, les processus « habituels/pavloviens » dépendent de l'automatisation progressive des interactions expertes ou des tendances innés d'approches ou d'évitements, lesquelles possèdent des propriétés particulières. Dans cette perspective, les interactions expertes et pavloviennes sont couplées à l'environnement pour permettre la minimisation de l'énergie libre variationnelle en employant des valeurs précédemment calculées ou « mises en cache » ou des valeurs spécifiés par des priors adaptatifs. Les premières sont

apprises en observant des séquences d'action « dirigées vers des buts » dans un contexte particulier tandis que les secondes sont des tendances comportementales innées qui sont spécifiées par l'évolution. Pour y parvenir, les processus « habituels/pavloviens » génèrent de boucles action-perception couplées avec l'environnement qui minimisent directement l'énergie libre variationnelle durant l'action experte ou pavlovienne (qui se déroule sans planification). En ce sens, nous pouvons caractériser la notion d'interaction experte et pavlovienne dans la perspective d'un processus d'ajustement dynamique « online » qui repose sur des modèles génératifs continus. En effet, les modèles génératifs continus permettent l'énaction automatique des séquences d'action en fonction des valeurs « mises en cache » ou de priors adaptatifs, sans demander la manipulation représentationnelle de croyances « à propos » des états futurs et conséquences associées. Ce faisant, les processus « habituels/pavloviens » peuvent être compris comme impliquant des interactions expertes ou pavloviennes étant « dirigées vers » une « emprise optimale sur le champ d'affordances » (Bruineberg & Rietveld, 2014) puisqu'ils impliquent l'énaction automatique de séquences d'actions « mises en cache » ou spécifiées de manière innée en fonction d'observations ou d'états cachés actuels. Les processus « habituels/pavloviens » n'impliquent pas de représentations avec des « conditions de satisfaction » puisque les croyances « à propos » des états cachés et conséquences préférées de l'action sont remplacées par des actions directes (association états-action).

D'un point de vue neurobiologique, les processus « habituels » vont engager des boucles thalamocorticales sensorimotrices avec des projections vers le striatum dorsolatéral et le cervelet tandis que les processus « pavloviens » vont engager des boucles thalamocorticales sensorimotrices avec des projections vers le striatum ventral et l'amygdale (Everitt & Robbins, 2013; Pezzulo, Rigoli et Friston, 2015). Les boucles thalamocorticales sensorimotrices utilisent les représentations continues de la voie du « où/comment » pour générer des « dynamiques sensorimotrices » (Buzsaki & Friston, 2016). Les « dynamiques sensorimotrices » sont des patrons d'interactions dynamiques métastables qui sont sélectivement réactifs aux affordances (en fonction des habiletés de l'agent). Dans la perspective de la MRH, cette conception des processus « habituels » correspond au processus de compétition biaisée d'affordance dans lequel des trajectoires d'actions alternatives compétitionnent jusqu'à ce que l'action optimale soit sélectionnée, cela sans l'intervention de modules centraux de croyances et de désirs. Comme nous l'avons vu dans le premier chapitre, la compétition d'affordances peut être comprise comme un processus d'ajustement dynamique qui permet de réduire la tension ressentie (ou minimiser l'énergie libre variationnelle) dans des situations familières en étant sélectivement réactif aux opportunités d'actions

dans l'environnement. Dans cette perspective, les buts, compris en termes de croyances et désirs, ne sont pas au fondement des actions expertes et pavloviennes, c'est plutôt la modulation continue d'un système dynamiquement couplé qui permet des interactions sensorimotrices adaptées au contexte.

Nous pouvons maintenant mieux comprendre en quoi le dynamicisme non-représentationnel et le computationnalisme représentationnel sont complémentaires dans la perspective d'un pluralisme explicatif qui permet l'intégration du fonctionnalisme et du structuralisme. Comme nous venons de l'aborder, les processus « dirigés vers des buts » impliquent la manipulation de représentations découplées de l'environnement tandis que les processus « habituels/pavloviens » impliquent la gestion des interactions couplées avec l'environnement. Du point de vue du fonctionnalisme, les processus « dirigés vers des buts » reposent sur la manipulation de représentations dans le système nerveux pour sélectionner l'action optimale (sélection par l'agent ou « goal-driven »). Il s'agit d'une explication fonctionnelle puisque ce sont les composantes constitutives qui déterminent la fonction. Du point de vue du structuralisme, les processus « habituels/pavloviens » reposent sur la gestion des interactions dans le système cerveau-corps-environnement pour sélectionner l'action optimale (sélection par l'environnement ou « stimulus-driven »). Il s'agit d'une explication structurelle puisque ce sont les interactions systémiques qui déterminent la fonction. En somme, l'auto-organisation globale « bottom-up » des processus « habituels/pavloviens » permet l'émergence d'une sélection locale « top-down » des processus « dirigés vers des buts », lesquels internalisent le contrôle de l'action dans des représentations dynamiques des états futurs ou buts.

2.2.2 L'architecture cognitive modulaire et interactive sous l'esprit hiérarchiquement mécaniste

Dans le premier chapitre, nous avons vu que l'hypothèse de modularité massive (MMH) décrit les fonctions cognitives comme émergeant de l'interaction entre modules fonctionnellement spécialisés dont les opérations internes sont largement inaccessibles aux autres sous-systèmes (c'est-à-dire par encapsulation dite « wide-scope »). Ces modules « domaine-spécifiques » sont des sous-systèmes fonctionnellement dissociables avec une réalisation neuronale spécifique (bien que parfois distribuée sur plusieurs régions neuronales) (association structure-fonction « one-to-one ») (Carruthers, 2006; 2008). Par ailleurs, nous avons aussi vu que l'hypothèse du redéploiement massif (MRH) décrit les fonctions cognitives comme émergeant de l'assemblage transitoire de sous-systèmes neuronaux locaux (TALoNS) dont les biais fonctionnels locaux sont contraints par des interactions globales. La contribution fonctionnelle des régions fonctionnellement différenciées est analysée en identifiant des ensembles de facteurs

neuroscientifiquement pertinents (NRP) « domaine-généraux » dans des empreintes fonctionnelles multidimensionnelles (association structure-fonction « many-to-many ») (Anderson, 2014; 2016). Dans cette section, nous discutons de la manière dont l'esprit hiérarchiquement mécaniste (HMM) pourrait permettre d'envisager la complémentarité entre ces deux hypothèses en distinguant entre mécanismes neurocognitifs « domaine-spécifiques » permettant la ségrégation locale modulaire aux niveaux inférieurs et mécanismes neurocognitifs « domaine-généraux » permettant l'intégration globale intermodulaire aux niveaux supérieurs.

Selon notre proposition, les mécanismes neurocognitifs « domaine-spécifiques » localement ségrégués sont associée à des modules fonctionnellement spécialisés, tandis que les mécanismes neurocognitifs « domaine-généraux » globalement intégrés sont associés aux hubs « rich-club » fonctionnellement différenciés, ceux-ci permettant la reconfiguration en temps réel de l'interaction neuronale nécessaire pour une tâche domaine-générale. Dans la perspective d'un pluralisme explicatif intégratif, nous soutiendrons que l'intégration fonctionnelle est issue d'un processus d'auto-organisation globale (explication structuraliste) tandis que la ségrégation fonctionnelle est issue d'un processus de sélection locale de composantes fonctionnelles (explication fonctionnaliste). Sous le HMM, les mécanismes neurocognitifs « domaine-spécifiques » de niveaux inférieurs sont contextualisés par les mécanismes neurocognitifs « domaine-généraux » de niveaux supérieurs en situation d'incertitude contextuelle (inférences profondes). Plus précisément, nous soutiendrons que les mécanismes neurocognitifs « domaine-spécifiques » (c'est-à-dire les modules avec des connexions de courte portée denses) aux niveaux inférieurs sont hiérarchiquement redéployés dans « l'espace de travail global » sous-tendu par des mécanismes neurocognitifs « domaine-généraux » (c'est-à-dire les hubs connecteurs « rich-club » avec des connexions de longue portée éparses) aux niveaux supérieurs en présence d'incertitude contextuelle pour permettre les processus délibérés ou « dirigés vers des buts ».

Dans la HMM, la ségrégation fonctionnelle est une conséquence de la factorisation des représentations des causes des sensations dans diverses « approximations du champ moléculaire » (mean-field approximation, Friston & Buzsaki, 2016). En des termes plus simples, la factorisation permet de décomposer une quantité qu'on comprend comme étant le produit de divers facteurs qu'on peut ensuite re-multiplier pour reproduire la quantité originale. Les « approximations du champ moléculaire » permettent ainsi une approximation des dépendances entre plusieurs facteurs pour permettre un produit plus « tractable » qui facilite l'encodage et l'actualisation des croyances (Friston & Buzsaki, 2016). Chacune

des factorisations possibles représentent différentes « approximations du champ moléculaire » ou différentes façons de produire un découpage statistique approximatif de « la nature à ses joints ». Dans la perspective de la neuroanatomie fonctionnelle, ce découpage statistique (factorisation) correspond à la ségrégation fonctionnelle, c'est-à-dire que les facteurs vont représenter les causes cachées statistiquement ou conditionnellement indépendantes des sensations qui correspondent aux attributs qui définissent la spécialisation fonctionnelle (par exemple, la couleur, le mouvement, les formes, etc.). En d'autres termes, la sélection naturelle, les facteurs épigénétiques et la plasticité dépendante de l'activité fournissent une « approximation du champ moléculaire » qui permet d'inférer les facteurs qui mènent aux sensations. Par exemple, connaître la couleur d'un objet n'indique (généralement) rien sur son mouvement, ce qui indique la présence d'une séparation anatomique dans les réseaux du cerveau. Dans les réseaux du cerveau, l'indépendance conditionnelle permet d'identifier les états cachés factorisés (ou modularisés) par une absence de connexion entre régions neuronales, ce qui permet la ségrégation de fonctions spécialisées lorsqu'il y a peu de connexions extrinsèques (voir section 2.1.2). En résumé, « conditional independence means that knowing the activity of one area tells you nothing about the activity of a second area, given the activity in all other areas. If this can be shown statistically, one can infer the absence of a connection between the two areas in question » (Friston & Price, 2011, p. 242).

Dans la HMM, l'intégration fonctionnelle dépend de la transmission de messages par des interactions dynamiquement couplées entre régions du cerveau (facteurs qui encodent des états cachés). La transmission de messages s'effectue par codage prédictif généralisé pour les états continus et, pour les états discrets, par la « propagation des croyances » (Friston, Parr et de Vries, 2017), ce qu'on peut formuler en termes de dynamiques neuronales comme une descente de gradient sur l'énergie libre variationnelle (Friston, Parr et de Vries, 2017). En d'autres termes, les dynamiques neuronales impliqueraient une transmission biconditionnelle hiérarchique de messages entre représentations factorisées, ce qui peut être formalisé dans la théorie des systèmes dynamiques en termes de descente de gradient sur l'énergie libre. Cette transmission hiérarchique bidirectionnelle de messages repose sur l'échange de statistiques suffisantes entre régions neuronales ségréguées (ou états cachés factorisés). Les statistiques suffisantes qui encodent des croyances postérieures (ou actualisées) sur les choses qui doivent être inférées. En effet, les statistiques suffisantes d'un modèle génératif sont des paramètres qui encodent uniquement les croyances postérieures (c'est-à-dire les croyances actualisées par les évidences), ce qui peut être interprété comme la probabilité attendue d'être dans un état particulier ou de poursuivre une séquence d'action particulière.

Selon notre proposition, les mécanismes neurocognitifs « domaine-spécifiques » peuvent être interprétés du point de vue de la modularité quasi-décomposable de la MMH. Nous soutenons que la ségrégation locale dans des modules par des connexions denses de courtes portées peut être interprétée dans la perspective de modules fonctionnellement spécialisés, dont les opérations internes sont largement inaccessibles aux autres sous-systèmes (encapsulation « wide-scope ») (Carruthers, 2006). Comme nous l'avons abordé dans le premier chapitre, les modules computationnels sont caractérisés par la notion de « dominance componentielle » qui soutient que les interactions dans les composantes sont plus importantes que les interactions entre les composantes (quasi-décomposabilité) (Van Orden, Holden et Turvey, 2003). Nous proposons d'envisager la dominance componentielle des modules spécialisés comme une conséquence de la ségrégation locale des représentations dans des facteurs (ou modules) conditionnellement indépendants qui peuvent être optimisés (ou modifiés) séparément. En ce sens, les opérations internes des modules factorisés sont largement inaccessibles aux autres facteurs, mais ceux-ci ne sont pas complètement indépendants puisqu'ils demeurent connectés aux autres facteurs par leur « champ moléculaires » (Parr, Sajid et Friston, 2020). Puisque seules les statistiques suffisantes sont échangées entre les modules factorisés, les distributions de probabilités complètes (ou croyances) ne sont pas elles-mêmes propagées ou échangées entre les niveaux hiérarchiques ou modules factorisés. Comme nous l'avons précédemment abordé, les modules factorisés sont associés à des régions neuronales anatomiquement ségréguées qui sont caractérisés par des connexions denses de courte portée qui permettent des dynamiques d'intégration locales rapides, mais restent connectés au reste du réseau par des connexions éparses de longue portée qui permettent des dynamiques d'intégration globales plus lentes (Friston & Buzsaki, 2016). En ce sens, nous pouvons caractériser les modules factorisés comme des sous-systèmes quasi-décomposables puisque les interactions dans les composantes sont plus importantes que les interactions entre les composantes.

Du point de vue des réseaux neuronaux du cerveau, les mécanismes neurocognitifs « domaine-spécifiques » vont correspondre aux modules spécialisés de niveaux inférieurs qui reposent sur des connexions denses de courte portée. Les modules permettent des régions neuronales spécialisées avec des fonctions spécifiques qui peuvent opérer de façon relativement autonome dans certains contextes. Nous soutenons que la ségrégation fonctionnelle locale implique une convergence structure-fonction (« one-to-one ») qui est essentielle pour les processus automatiques « habituels/pavloviens ». En effet, la convergence structure-fonction est principalement observable dans des réseaux de connectivité intrinsèque (intrinsic connectivity networks ou ICN) qui regroupent des réseaux tels que le réseau en mode

par défaut, le réseau d'attention dorsale, le réseau de saillance, les réseaux sensori-moteurs, etc. (Park & Friston, 2013). Les ICNs sont décomposables en sous-modules composés de nœuds en fluctuations synchrones persistantes qui sont médiées par leurs interconnexions denses de courte portée. Les analyses de connectivité révèlent une correspondance considérable entre les ICNs et des modules neurocognitifs associés à des tâches (Park & Friston, 2013; voir aussi Bertolero, Yeo et d'Esposito, 2015). Dans cette perspective, la connectivité fonctionnelle des sous-modules des ICNs ressemble fortement à la connectivité structurelle sous-jacente, ce qui indique que l'organisation hiérarchiquement modulaire et « rich-club » des réseaux fournit des contraintes structurelles lors de l'intégration fonctionnelle « sensible au contexte » (Park & Friston, 2013). Ces contraintes structurelles qui favorisent des états hautement ségrégués des réseaux neuronaux peut être comprise dans la perspective des processus heuristiques ou « habituels/pavloviens », c'est-à-dire des processus qui n'impliquent pas d'intégration globale « sensible au contexte » qui repose sur la mémoire de travail (Cohen & D'Esposito, 2016; Bassett et al., 2015). Dans la perspective de la MMH, nous pouvons envisager la convergence structure-fonction comme une conséquence d'une encapsulation « wide-scope » qui repose sur des heuristiques de recherche et des règles d'arrêt. En effet, lorsque l'incertitude contextuelle est faible, les processus automatiques ou « habituels/pavloviens » minimisent l'énergie libre localement dans des modules spécialisés, sans recruter « l'espace de travail global » avec ses connexions de longues portées métaboliquement coûteuses. De cette façon, les processus automatiques peuvent être réalisés par des modules avec une synchronisation dynamique locale qui permet d'accomplir une tâche sans délibération consciente (voir Dehaene, Changeux et Naccache, 2011).

Selon notre proposition, les mécanismes neurocognitifs « domaines-généraux » peuvent être interprétés du point de vue de l'interactivité non-décomposable de la MRH. Nous soutenons que l'intégration globale dans des hubs « rich-club » par des connexions de longues portées éparses peut être interprétée dans la perspective d'assemblages transitoires de sous-systèmes neuronaux locaux (TALoNS) dont les biais fonctionnels locaux sont contraints par des interactions globales (Anderson, 2014). Comme nous l'avons abordé dans le premier chapitre, les interactions dynamiques sont caractérisées par une « dominance interactive » qui soutient que les interactions entre les composantes sont plus importantes que les interactions dans les composantes (non-décomposabilité) (Orden, Holden et Turvey, 2003). Nous proposons d'envisager la dominance interactive des TALoNS comme une conséquence de l'intégration globale par la transmission hiérarchique de messages entre différents facteurs conditionnellement indépendants, ce qui permet de former dynamiquement des coalitions fonctionnelles dans des ensembles

coordonnés multifactoriels. En ce sens, les interactions globales non-linéaires dans des hubs « rich-club » permettent de coordonner la formation de différentes coalitions fonctionnelles multifactorielles avec des propriétés interactives non-décomposables. Puisque les dynamiques neuronales globales reposent sur l'échange bidirectionnel de statistiques suffisantes dans les réseaux hiérarchiques du cerveau, elles réalisent des propriétés interactives émergentes qui diffèrent de celles des composantes fonctionnelles qui les constituent. En effet, les propriétés fonctionnelles globales qui émergent des interactions non-linéaires multifactorielles dans des hubs « rich-club » massivement interconnectés peuvent être réalisées par différentes combinaisons de modules factorisés (réalisation multiple des propriétés systémiques qui émergent des interactions non-linéaires entre composantes). Ce faisant, une décomposition mécaniste des interactions multifactorielles globales éliminerait d'emblée une explication de ces propriétés fonctionnelles globales qui émergent d'interactions non-linéaires. En ce sens, nous pouvons caractériser les coalitions fonctionnelles multifactorielles comme des interactions non-décomposables puisque les interactions entre les composantes sont plus importantes que les interactions dans les composantes (lorsque vient le temps de comprendre des propriétés systémiques comme la conscience, la mémoire de travail, l'attention, etc.).

Du point de vue des réseaux neuronaux du cerveau, les mécanismes neurocognitifs « domaine-généraux » de niveaux supérieurs vont correspondre aux hub connecteurs « rich-club » qui reposent sur des connexions éparses de longue portée. Les hubs connecteurs « rich-club » multifonctionnels vont permettre le redéploiement des modules dans des coalitions fonctionnelles multifactorielles « sensibles au contexte ». Nous soutenons que l'intégration fonctionnelle globale implique une divergence structure-fonction (« many-to-many ») qui est essentielle pour les processus « dirigés vers des buts ». En effet, la divergence structure-fonction est principalement observable durant l'intégration globale « sensible au contexte » (Park & Friston, 2013). Cette intégration globale par des connexions à longue portée repose sur la modulation du gain synaptique (c'est-à-dire de la pondération de la précision) et/ou des interactions non-linéaires synchronisées (c'est-à-dire du couplage dynamique) qui permettent une auto-organisation globale de l'architecture du réseau. C'est cette particularité des réseaux neuronaux qui permet de générer plusieurs patrons de connectivité fonctionnelle sur une connectivité structurelle relativement fixe. Les connexions longue portée vont permettre une intégration intermodulaire « sensible au contexte » grâce à la reconfiguration dynamique du réseau fonctionnel, en recrutant divers assemblages de modules spécialisés dans des patrons d'activité métastables qui émergent de l'intégration dynamique non-linéaire. Étant donné que la connectivité effective est considérablement contrainte par la connectivité structurelle,

l'intégration fonctionnelle (par le biais de la transmission de messages neuronaux) peut alors servir d'interface entre la structure et la fonction, ce qui expliquerait les associations divergentes « many-to-many » observées (Park & Friston, 2013). Dans la perspective de la MRH, nous pouvons ainsi envisager la divergence structure-fonction comme une conséquence d'un processus de recherche neuronal qui repose sur l'essai rapide d'une multitude de combinaisons neuronales pour trouver la plus appropriée lors de l'acquisition d'une habileté (Anderson 2014). En effet, lorsque l'incertitude contextuelle est importante, les processus délibérés (« dirigés vers des buts ») minimisent l'énergie libre par synchronisation dynamique globale dans « l'espace de travail global » avec ses connections de longues portées qui permettent de recruter différentes configurations de modules avant de trouver la configuration appropriée pour la tâche. Ce faisant, les interactions non-linéaires dans « l'espace de travail » massivement interconnecté vont aussi entraîner une « ignition » qui recrute les connexions de longues portées, laquelle est nécessaire pour permettre l'émergence de la délibération consciente (voir Dehaene, Changeux et Naccache, 2011).

Nous pouvons maintenant mieux comprendre en quoi la modularité quasi-décomposable et l'interactivité non-décomposable sont complémentaires dans la perspective d'un pluralisme explicatif qui permet l'intégration du fonctionnalisme et du structuralisme. Comme nous venons de le voir, la ségrégation locale dans des modules spécialisés implique la quasi-décomposabilité tandis que l'intégration globale dans des hubs connecteurs « rich-club » multifonctionnels implique la non-décomposabilité. Du point de vue du fonctionnalisme, la ségrégation locale dans des modules quasi-décomposables permet d'implémenter des fonctions spécialisées par leurs connexions denses de courtes portées. Il s'agit d'une explication fonctionnelle puisque ce sont les interactions causales entre composantes constitutives dans des modules qui déterminent la fonction (explication componentielle mécaniste). Du point de vue du structuralisme, l'intégration globale dans des hubs « rich-club » permet d'implémenter des coalitions multifonctionnelles par leurs connexions éparses de longues portées. Il s'agit d'une explication structurelle puisque ce sont les interactions systémiques dans des hubs « rich-club » qui déterminent la fonction (explication non-componentielle par contrainte habilitante). En somme, l'auto-organisation globale « bottom-up » des modules spécialisés permet l'émergence d'une sélection locale « top-down » par des hubs connecteurs « rich-club », lesquels internalisent le contrôle de l'activité neuronale dans des séquences hippocampiques générées à l'interne qui permettent la planification « dirigés vers des buts ».

2.2.3 L'architecture cognitive adaptée et adaptative sous la théorie des systèmes évolutionnaires

Dans le premier chapitre, nous avons vu que l'hypothèse de la modularité massive (MMH) décrit une architecture fonctionnellement spécialisée comme étant façonnée par des pressions sélectives locales issues d'un processus de sélection naturelle. Le neurodéveloppement canalisé repose sur l'interaction entre le programme développemental et les caractéristiques environnementales typiquement récurrentes pour générer les modules (d'apprentissage) « domaine-spécifiques » (Carruthers, 2006). Par ailleurs, nous avons aussi vu que l'hypothèse de la réutilisation neuronale (MRH) décrit une architecture fonctionnellement différenciée comme étant façonnée par des pressions sélectives globales issues d'un processus d'auto-organisation évo-dévo. Le neurodéveloppement plastique reposerait selon Anderson (2014) sur un processus de recherche et différentiation interactive (IDS) qui dépend de l'activité pour générer les charges sur des biais fonctionnels « domaine-généraux ». Dans cette section, nous discutons de la manière dont la théorie des systèmes évolutionnaires (EST, Badcock 2012) permet d'envisager la complémentarité entre ces deux hypothèses en distinguant entre pressions sélectives locales qui favorisent l'évolution mosaïque de modules spécialisés et pressions sélectives globales qui favorisent l'évolution concertée de hubs connecteurs « rich-club ».

Selon notre proposition, les pressions sélectives locales qui reposent sur des contraintes fonctionnelles génèrent des mécanismes neurocognitifs « domaine-spécifiques » significativement canalisés aux niveaux inférieurs de l'architecture cognitive et les pressions sélectives globales qui reposent sur des contraintes développementales génèrent des mécanismes neurocognitifs « domaine-généraux » significativement plastiques à ses niveaux supérieurs. Dans la perspective d'un pluralisme explicatif intégratif, nous soutiendrons que les mécanismes neurocognitifs « domaine-spécifiques » sont principalement issus d'un processus de sélection locale de composantes fonctionnelles (et donc sujets à une explication fonctionnaliste) tandis que les mécanismes neurocognitifs « domaine-généraux » sont principalement issus d'un processus d'auto-organisation global (et donc sujets à une explication structuraliste). Sous la théorie des systèmes évolutionnaires (EST), le phénotype est façonné par l'interaction dynamique entre la sélection et l'auto-organisation sur les quatre niveaux d'analyse biologique proposés par Tinbergen (1963). Plus précisément, la sélection naturelle s'exerce principalement sur les niveaux « sensorimoteurs » pour permettre un neurodéveloppement canalisé de modules et l'auto-organisation s'exerce principalement sur les niveaux « associatifs » pour permettre un neurodéveloppement plastique de hubs connecteurs « rich-club ».

Dans la EST, le niveau d'analyse de l'adaptation (des fonctions) est associé à la psychologie évolutionniste de la synthèse moderne néo-darwinienne qui soutient une conception sélectionniste de l'évolution qui repose d'abord sur la sélection naturelle. Ce niveau d'analyse est associé à l'échelle temporelle évolutionnaire, laquelle permet d'optimiser l'énergie libre moyenne des membres d'une espèce par la transmission génétique des priors adaptatifs qui supportent les processus d'apprentissage pavloviens, notamment. La psychologie évolutionniste de la synthèse moderne implique des pressions sélectives locales qui concernent des problèmes adaptatifs dans l'environnement évolutionnaire. En effet, ce paradigme explore comment les processus évolutionnaires génèrent les adaptations (fonctions) qui sous-tendent les comportements humains. Les adaptations sont typiquement comprises comme des traits adaptatifs qui ont évolué par sélection naturelle pour accroître les chances de survie et de reproduction, ce qui est compris en termes de fitness. La fitness mesure la capacité de l'organisme à surmonter des problèmes adaptatifs, ce qui permet à l'organisme de transmettre les gènes qui encodent les traits adaptatifs aux générations subséquentes. Dans cette perspective, la transmission génétique permet des adaptations qui répondent aux problèmes adaptatifs récurrents dans l'environnement évolutionnaire. Dans la EST, la transmission génétique permet de générer des mécanismes neurocognitifs « domaine-spécifiques » qui sont suffisamment canalisés pour permettre des réponses spécifiquement adaptées à l'environnement évolutionnaire (mais peut-être pas à des environnements différents).

Dans la EST, le niveau d'analyse de la phylogénie (évolution) (qui concerne ici les changements phylogéniques sur une échelle temporelle intergénérationnelle) est associé à la biologie/psychologie évo-dévo de la synthèse évolutionnaire étendue qui soutient une conception développementaliste de l'évolution qui repose d'abord sur l'auto-organisation. Ce niveau d'analyse est associé à l'échelle temporelle intergénérationnelle, laquelle permet d'optimiser l'énergie libre moyennes des membres d'un sous-groupe appartenant à une espèce par la transmission épigénétique/exogénétique des priors adaptatifs qui supportent les processus d'apprentissage instrumentaux, notamment. La biologie/psychologie évo-dévo de la synthèse étendue implique des pressions sélectives (globales) qui concernent la physiologie de l'organisme. En effet, ce paradigme explore les dynamiques par lesquelles les changements développementaux (ontogénique) interagissent avec les changements entre les générations (phylogénie). Une forme importante de transmission des variations ontogéniques est l'épigénétique, laquelle réfère à la transmission de variations phénotypiques qui ne nécessite pas des altérations dans le génome, ce qui permet de fournir des nouvelles cibles pour la sélection naturelle (Jablonka & Lamb, 2005). Une autre forme importante de transmission des variations ontogéniques est la

transmission éxogénétique, laquelle réfère à la transmission des ressources environnementales qui sont nécessaires aux cycles de reproduction des organismes, ce qui inclut la construction de niche et la transmission intergénérationnelle d'information culturelle accumulée (Sterelny, 2003). Dans cette perspective, la capacité de transmission épigénétique/exogénétique constitue elle-même une adaptation qui permet de compenser les limitations de la transmission génétique en permettant une transmission intergénérationnelle qui permet de compenser relativement rapidement pour les changements environnementaux. Dans la EST, la transmission épigénétique/exogénétique permet de générer des mécanismes neurocognitifs « domaine-généraux » qui sont suffisamment plastiques pour permettre des réponses flexibles qui s'adaptent aux environnements changeants.

Selon notre proposition, les mécanismes neurocognitifs « domaine-spécifiques » aux niveaux inférieurs peuvent être interprétés du point de vue de l'adaptation de la MMH. Nous soutenons que les pressions sélectives locales façonnent des modules spécialisés avec des connexions denses de courte portée. Comme nous l'avons abordé dans le premier chapitre, l'architecture fonctionnellement spécialisée peut être comprise dans la perspective d'une évolution mosaïque qui est caractérisée par des pressions sélectives locales pour la résolution de problèmes adaptatifs dans l'environnement évolutionnaire. Nous proposons d'envisager les pressions sélectives locales comme étant essentielle pour l'évolution mosaïque de mécanismes « domaine-spécifiques » significativement ségrégués qui reflètent des régularités statistiques préservées sur une échelle temporelle évolutionnaire. En effet, l'évolution mosaïque dépend de contraintes fonctionnelles qui permettent l'évolution de régions neuronales de manière relativement indépendantes. Les contraintes fonctionnelles agissent sur des composantes fonctionnellement interconnectées pour réaliser des fonctions spécialisées (Barton & Harvey, 2000). Dans cette perspective, les contraintes fonctionnelles reposent sur la modifiabilité séparée des composantes, ce qui est essentiel pour une ségrégation efficace des fonctions cognitives dans des modules factorisés. En effet, les pressions locales concernent des problèmes adaptatifs externes à l'organisme qui sont résolus en façonnant des modules fonctionnellement spécialisés. Les modules permettent de résoudre les problèmes adaptatifs externes en permettant des solutions spécialisées qui assurent la survie et la reproduction. Pour y parvenir, les modules utilisent des connexions denses de courte portée qui assurent la spécialisation fonctionnelle de facteurs ségrégués, mais une multiplication des modules peut entraîner des problèmes d'efficacité global de traitement et de coût métabolique. En effet, les modules favorisent l'efficacité local de traitement, mais l'ajout incrémentiel de modules risque de perturber le fonctionnement des modules préexistants en plus d'encourir des coûts métaboliques (voir Sporn, 2011).

D'un point de vue neurodéveloppemental, les mécanismes neurocognitifs « domaine-spécifiques » des niveaux inférieurs sont significativement canalisés. Nous pouvons associer la canalisation à une forme de robustesse conférée par la modularité du système. La modularité est une propriété locale qui permet de conserver la fonctionnalité du système en permettant aux perturbations d'affecter des composantes sans affecter l'ensemble du système (modifiabilité séparée). Nous suggérons d'envisager la canalisation comme une forme de robustesse qui est réalisée sur l'échelle temporelle évolutionnaire. Dans cette perspective, la sélection naturelle permet au génotype d'une espèce de conserver un fitness constant, malgré des fluctuations environnementales. En d'autres termes, la relation entre génotype et phénotype est maintenue constante de façon à préserver le fitness. Dans la perspective de la MMH, cette conception de la canalisation neurodéveloppementale peut être associée au programme développemental. Selon le programme développemental, un développement fonctionnel robuste émerge d'interrupteurs génétiques qui activent ou désactivent le développement de modules spécialisés (Carruthers, 2006). Nous pouvons comprendre le neurodéveloppement des modules spécialisés dans la perspective du programme développemental puisqu'ils sont des mécanismes « domaine-spécifiques » qui sont façonnés par des interrupteurs génétiques qui contrôlent l'expression génétique pour façonner des sous-systèmes locaux.

Selon notre proposition, les mécanismes neurocognitifs « domaine-généraux » aux niveaux supérieurs peuvent être interprétés du point de vue de l'adaptativité de la MRH. Nous soutenons que les pressions sélectives globales façonnent des hubs connecteurs « rich-club » avec des connexions éparses de longue portée. Comme nous l'avons abordé dans le premier chapitre, l'architecture fonctionnellement différenciée peut être comprise dans la perspective d'une évolution concertée qui est caractérisée par des pressions sélectives globales pour la résolution de problèmes physiologiques et informationnels dans l'organisme. Nous proposons d'envisager les pressions sélectives globales comme étant essentielle pour l'évolution concertée des mécanismes neurocognitifs « domaine-généraux » significativement intégrés qui reflètent des régularités statistiques sur une échelle temporelle intergénérationnelle. En effet, l'évolution concertée dépend de contraintes développementales qui permettent l'évolution du cerveau comme un ensemble intégré. Les contraintes développementales agissent sur les processus développementaux qui affectent la croissance de l'ensemble des composantes pour limiter les phénotypes possibles, mais aussi permettre aux changements phylogénétiques de se produire dans certaines directions plutôt que d'autres (Finlay & Darlington, 1995). Dans cette perspective, les contraintes développementales reposent sur l'efficacité métabolique et l'efficacité de traitement globale dans l'architecture cognitive, ce qui est essentiel pour une intégration efficace des fonctions cognitives dans des assemblages multifactoriels. En

effet, les pressions globales concernent les problèmes physiologiques et informationnels internes à l'organisme qui sont résolus en façonnant des hubs connecteurs « rich-club » fonctionnellement différenciés. Les hubs « rich-club » permettent de résoudre les problèmes physiologiques et informationnels internes en permettant une coordination efficace de l'intégration fonctionnelle intermodulaire (efficacité de traitement globale) et en minimisant le coût métabolique entraîné par l'ajout incrémentiel de modules spécialisés (efficacité métabolique). Pour y parvenir, les hubs connecteurs « rich-club » utilisent des connexions éparses de longue portée qui assurent le redéploiement de ressources neuronales pré-existantes dans des assemblages multifactoriels fonctionnellement intégrés. En effet, les hubs connecteurs « rich-club » sont essentiels pour permettre l'efficacité de traitement globale (c'est-à-dire l'intégration fonctionnelle) tout en minimisant les coûts de connexions, cela malgré le coût métabolique des connexions de longue portée.

D'un point de vue neurodéveloppemental, les mécanismes neurocognitifs « domaine-généraux » sont significativement plastiques. Nous pouvons associer la plasticité à une forme de robustesse conférée par la complexité. La complexité est une propriété globale qui permet de conserver la fonctionnalité du système puisqu'elle repose sur différentes configurations de composantes qui peuvent compenser les perturbations de plusieurs façons différentes (c'est-à-dire en permettent une réalisation multiple des propriétés systémiques). En effet, la plasticité phénotypique est une forme de robustesse qui est réalisée sur l'échelle temporelle intergénérationnelle. Dans cette perspective, l'auto-organisation permet de maintenir un fitness constant par des changements phénotypiques sensibles aux variations environnementales, sans demander de modification dans le génotype. En d'autres termes, la relation entre le génotype et le fitness demeure constante au détriment de la relation entre le génotype et le phénotype qui devient changeante pour permettre plusieurs phénotypes à partir d'un seul génotype. Dans la perspective de la MRH, cette conception de la plasticité neurodéveloppementale peut être associée au processus de recherche et différenciation interactive (IDS) (Anderson, 2014). Selon l'IDS, le développement fonctionnel flexible émerge de la différenciation interactive locale des biais fonctionnels « domaine-généraux » et d'un processus de recherche neuronale qui permet d'identifier les partenariats fonctionnels adéquats. Nous pouvons comprendre le neurodéveloppement des hubs connecteurs « rich-club » dans la perspective de l'IDS puisqu'ils sont des mécanismes neurocognitifs « domaine-généraux » qui acquiert leurs fonctions par l'interaction globale dépendante de l'activité. En situation d'incertitude contextuelle, les hubs connecteurs « rich-club » s'activent pour permettre la recherche des partenariats

fonctionnels dans tout le cerveau, lesquels vont former les assemblages transitoires de sous-systèmes neuronaux locaux (TALoNS).

Nous pouvons maintenant mieux comprendre comment l'adaptation de l'évolution mosaïque et l'adaptativité de l'évolution concertée sont complémentaire dans la perspective d'un pluralisme explicatif qui permet l'intégration du fonctionnalisme et du structuralisme. Comme nous venons de l'expliquer, l'évolution mosaïque repose sur des contraintes fonctionnelles locales qui opèrent sur une échelle temporelle évolutionnaire tandis que l'évolution concertée repose sur des contraintes développementales globales qui opèrent sur une échelle temporelle intergénérationnelle. Du point de vue du fonctionnalisme, les contraintes fonctionnelles de l'évolution mosaïque par sélection naturelle permettent de façonner des modules spécialisés sur une échelle temporelle évolutionnaire. Il s'agit d'une explication fonctionnelle puisque ce sont les pressions sélectives locales sur des composantes modulaires qui déterminent la fonction. Du point de vue du structuralisme, les contraintes développementales de l'évolution concertée par auto-organisation permettent de façonner des hubs connecteurs « rich-club » multifonctionnels sur une échelle temporelle intergénérationnelle. Il s'agit d'une explication structurelle puisque ce sont les pressions sélectives globales sur les processus développementaux qui déterminent la fonction. En somme, l'auto-organisation globale « bottom-up » par évolution concertée permet l'émergence d'une sélection locale « top-down » par l'évolution mosaïque qui internalise le contrôle de la plasticité neurodéveloppementale par assimilation génétique et canalisation.

2.3 Conclusion de chapitre

Dans ce chapitre, nous avons proposé d'envisager les thèses principales de la MMH et de la MRH comme étant complémentaire lorsqu'on les interprète dans la perspective d'un pluralisme explicatif intégratif sous le FEP (Friston, 2010). Nous avons commencé par introduire le FEP, un énoncé formel qui soutient que les systèmes biologiques opèrent selon un principe variationnel qui implique la minimisation d'une quantité informationnelle nommée énergie libre. Nous avons ensuite présenté la HMM, une théorie de l'architecture cognitive qui soutient que le cerveau est système adaptatif qui minimise l'énergie libre par des cycles action-perception qui émergent d'interactions dynamiques entre des mécanismes neurocognitifs différentiellement intégrés et ségrégués. La HMM incorpore aussi une EST qui soutient que l'optimisation des fonctions et structures s'effectuent par l'interaction entre la sélection (naturelle) et l'auto-organisation sur les échelles temporelles identifiées par les niveaux d'analyse de Tinbergen (1963), soit l'adaptation, la phylogénie, l'ontogénie et le mécanisme. Ensuite, nous avons soutenu que

l'optimisation du phénotype implique une interaction bidirectionnelle entre fonctionnalisme sélectionniste de la MMH et le structuralisme développementaliste de la MRH. Dans un premier temps, nous avons proposé d'envisager le computationalisme et le dynamicisme comme étant complémentaire dans l'inférence active profonde en usant d'une distinction entre processus « dirigés vers des buts » et processus « habituels/pavloviens ». Dans un second temps, nous avons proposé d'envisager la modularité et l'interactivité comme étant complémentaire dans la HMM en usant d'une distinction entre la ségrégation modulaire et intégration intermodulaire. Dans un troisième temps, nous avons proposé d'envisager l'adaptation et l'adaptativité comme étant complémentaire dans la EST en usant d'une distinction entre évolution mosaïque et évolution concertée. Dans la conclusion, nous proposons d'aborder les avantages d'une conception modulaire et interactive de l'architecture cognitive sous le FEP, notamment en discutant de la problématique de la conscience.

CONCLUSION

Dans ce mémoire, nous avons proposé que les thèses qui caractérisent l'hypothèse de la modularité massive (MMH) et l'hypothèse du redéploiement massif (MRH) entretiennent une relation complémentaire dans la perspective d'un pluralisme explicatif intégratif sous le principe d'énergie libre (FEP) (Friston, 2010). Selon notre interprétation, le FEP permettrait de dissoudre l'opposition entre la conception cognitiviste de la MMH, qui repose sur la manipulation de représentations découplées de l'environnement, et la conception incarnée de la cognition de la MRH, qui repose sur la gestion d'interactions couplées avec l'environnement. Tout au long du mémoire, nous soutenons que ces architectures cognitives s'attardent à des aspects différents, ce que nous proposons de comprendre dans la perspective d'une distinction entre fonctionnalisme biologique, qui procède des parties au tout, et structuralisme biologique, qui procède du tout aux parties.

Mon argument procède en deux moment : D'une part, nous avons soutenu que la MMH représente une application du fonctionnalisme sélectionniste à l'architecture cognitive. Cette proposition repose sur la synthèse moderne néo-darwinienne qui soutient que la sélection naturelle est la seule force capable d'expliquer l'évolution des systèmes fonctionnels complexes. Selon cette stratégie explicative, les organismes sont compris dans une perspective mécaniste-réductionniste comme des systèmes composites qui sont la somme de leurs composantes. En effet, la MMH soutient explicitement que l'architecture cognitive est composée de modules fonctionnellement spécialisés ayant évolué par processus de sélection naturelle. Dans le premier chapitre, nous avons soutenu que la MMH repose sur les trois thèses « componentielles » suivantes : (1) le computationnalisme représentationnel; (2) la modularité quasi-décomposable et; (3) l'adaptation par évolution mosaïque. Les thèses sont dites « componentielles » puisqu'elles reposent essentiellement sur des composantes quasi-décomposables dans un système nerveux. Puisque la cognition est alors vue comme un processus interne (internalisme) qui repose sur le système nerveux, elle est aussi comprise comme un processus computationnel de manipulation de représentations découplées de l'environnement. Cette conception de la cognition repose sur des modules computationnels qui sont des sous-systèmes quasi-décomposables dont les interactions à l'intérieur des composantes sont plus importantes que les interactions entre les composantes. Ce faisant, les pressions sélectives peuvent cibler des régions neuronales de façon quasi-indépendantes pour réaliser différentes fonctions cognitives, ce qu'on nomme l'hypothèse de l'évolution mosaïque de la cognition.

D'autre part, nous avons soutenu que la MRH représente une application du structuralisme développementaliste à l'architecture cognitive. Cette proposition repose sur la synthèse évolutionnaire étendue qui soutient que la sélection naturelle n'est pas la seule force capable d'expliquer l'évolution des systèmes fonctionnels complexes, mais que d'autres processus non-darwiniens comme l'auto-assemblage sont aussi nécessaires. Selon cette stratégie explicative qui procède de tout aux parties, les organismes sont compris dans une perspective organiciste-émergentiste comme des systèmes intégrés qui sont plus que la somme de leurs composantes. En effet, la MRH soutient explicitement que l'architecture cognitive est composée de régions fonctionnellement différenciées ayant évolué par processus évo-dévo. Dans le premier chapitre, nous avons soutenu que la NRH repose sur les trois thèses « processuelles » suivantes : (1) le dynamicisme non-représentationnel; (2) l'interactivité non-décomposable et; (3) l'adaptativité par évolution concertée. Les thèses sont dites « processuelles » puisqu'elles reposent essentiellement sur des processus non-décomposables dans un système cerveau-corps-environnement. Puisque la cognition est alors comprise comme un processus externaliste qui repose sur le système cerveau-corps-environnement, elle est comprise également comme un processus dynamique des interactions couplées avec l'environnement. Cette conception de la cognition repose sur des biais corticaux qui forment des « soft-assemblies » (Kello & Van Orden, 2003) non-décomposables dont les interactions entre les composantes sont plus importantes que les interactions dans les composantes. Ce faisant, les pressions sélectives peuvent cibler les processus développementaux pour permettre la disponibilité d'une vaste gamme de biais corticaux dans tout le cerveau, ce qu'on nomme l'hypothèse de l'évolution concertée de la cognition.

Pour mieux comprendre comment les thèses sont ultimement complémentaires dans un pluralisme explicatif intégratif, nous avons présenté une conception de l'architecture cognitive qui repose sur le FEP. Selon l'esprit hiérarchiquement mécaniste (HMM, Badcock et al., 2019), le cerveau est un système adaptatif complexe qui minimise l'énergie libre par des cycles action-perception générés par des dynamiques neuronales hiérarchiques bidirectionnelles entre mécanismes neurocognitifs différentiellement ségrégués et intégrés. Selon notre proposition, aux niveaux inférieurs se trouvent des mécanismes neurocognitifs « domaine-spécifiques » au neurodéveloppement canalisé qui sont localement ségrégués par des connexions denses de courtes portées. Aux niveaux supérieurs se trouvent des mécanismes neurocognitifs « domaine-généraux » au neurodéveloppement plastique qui sont globalement intégrés par des connexions éparses de longues portées. En somme, cette conception de l'architecture cognitive attribue une importance égale à la ségrégation locale et l'intégration globale : les mécanismes neurocognitifs sont relativement ségrégués et intégrés, ils possèdent tous des connexions

intrinsèques qui réalisent la ségrégation locale, mais possèdent aussi tous des connexions extrinsèques qui réalisent l'intégration globale.

La proposition avancée dans ce mémoire consistait à envisager la complémentarité entre les stratégies explicatives sous le FEP en prenant en compte la distinction entre les niveaux d'analyse biologique proposés par Tinbergen (1963). Plus particulièrement, nous distinguons dans une théorie des systèmes évolutionnaires (EST) entre les niveaux d'analyse biologique associés au fonctionnalisme sélectionniste, c'est-à-dire le mécanisme (causalité) et l'adaptation (fonction), et ceux associés au structuralisme développementaliste, c'est-à-dire le développement (ontogénie) et l'évolution (phylogénie). Dans la EST, la sélection naturelle et l'auto-organisation sont comprises comme des façons complémentaires de minimiser l'énergie libre du phénotype. Comme nous l'avons abordé dans le second chapitre, la distinction entre les niveaux d'analyse biologique s'opère d'abord en fonction des échelles temporelles considérées : (1) l'échelle temporelle évolutionnaire pour l'adaptation, qu'on associe à la psychologie évolutionniste et la synthèse moderne néo-darwinienne; (2) l'échelle temporelle intergénérationnelle, qu'on associé à la psychologie évo-dévo et à la synthèse évolutionnaire étendue; (3) l'échelle temporelle développementale, qu'on associe à la psychologie développementale et; (4) l'échelle temporelle en temps réel, qu'on associe aux diverses sous-disciplines de la psychologie. Dans cette perspective, la MMH sélectionniste s'attarde aux davantage niveaux d'analyse de l'adaptation (fonction) et du mécanisme (causalité) tandis que la MRH développementaliste s'attarde davantage aux niveaux d'analyse de l'évolution (phylogénie) et du développement (ontogénie). C'est pourquoi nous soutenons qu'il est possible d'envisager la complémentarité entre les explications fonctionnalistes proposées par la MMH et les explications structuralistes proposées par la MRH.

Sous l'inférence active profonde (Pezzulo et al., 2018), nous avons proposé que le computationnalisme représentationnel de la MMH et le dynamicisme non-représentationnel de la MRH concernent des aspects complémentaires dans l'architecture cognitive. Pour ce faire, nous avons proposé une distinction entre processus automatiques « habituels/pavloviens » aux niveaux inférieurs et processus délibérés « dirigés vers des buts » aux niveaux supérieurs. L'auto-organisation globale « bottom-up » des processus « habituels/pavloviens » dans le système cerveau-corps-environnement permet l'émergence de la sélection locale « top-down » des processus « dirigés vers des buts » qui contrôle la planification de l'action dans le système nerveux, ce qui démontre l'interaction bidirectionnelle entre le fonctionnalisme (parties-au-tout) et le structuralisme (tout-aux-parties).

Sous la HMM, nous avons proposé que la modularité quasi-décomposable de la MMH et l'interactivité non-décomposable de la MRH concernent des aspects complémentaires dans l'architecture cognitive. Pour ce faire, nous avons proposé une distinction entre ségrégation locale dans des modules « domaine-spécifiques » aux niveaux inférieurs et intégration globale dans des hubs « rich-club » « domaine-généraux » aux niveaux supérieurs. L'auto-organisation globale « bottom-up » des modules ségrégués dans le système cerveau-corps-environnement permet l'émergence de la sélection locale « top-down » des hubs connecteurs intégrés qui contrôle l'activité neuronale dans le système nerveux, ce qui démontre l'interaction bidirectionnelle entre le fonctionnalisme (parties-au-tout) et le structuralisme (tout-aux-parties).

Sous la EST, nous avons proposé que l'évolution mosaïque de la MMH et l'évolution concertée de la MRH concernent des aspects complémentaires dans l'architecture cognitive. Pour ce faire, nous avons proposé une distinction entre pressions sélectives locales pour l'évolution mosaïque des modules aux niveaux inférieurs et pressions sélectives globales pour l'évolution concertée des hubs connecteurs « rich-club » aux niveaux supérieurs. L'auto-organisation globale « bottom-up » des contraintes fonctionnelles dans le système cerveau-corps-environnement permet l'émergence de la sélection locale « top-down » des contraintes développementales qui contrôle l'expression génétique dans le système nerveux, ce qui démontre l'interaction bidirectionnelle entre le fonctionnalisme (parties-au-tout) et le structuralisme (tout-aux-parties).

Cette distinction entre les stratégies explicatives permet d'aborder certaines problématiques importantes en sciences cognitives sous un nouveau jour. En envisageant la complémentarité entre la perspective fonctionnaliste de la MMH, qui procède des parties au tout, et la perspective structuraliste de la MRH, qui procède du tout aux parties, nous pouvons mieux comprendre certaines propriétés fonctionnelles émergentes qui caractérisent l'architecture cognitive du cerveau. Par exemple, nous pouvons mieux étoffer des théories présentes dans la littérature, notamment la théorie de l'espace de travail neuronal global de la conscience d'accès (Dehaene, Changeux et Nacache, 2011). Nous avons brièvement abordé cette théorie, qui soutient que les boucles de répétition mentale consciente reposaient sur la diffusion globale dans un espace de travail global, dans le cadre de la MMH (Carruthers, 2006). En revanche, dans un souci de préserver une architecture entièrement composée de modules « domaine-spécifiques », Carruthers envisage l'espace de travail global comme un système virtuel qui émerge des interactions intermodulaires, ce qui diffère de la théorie originale de Dehaene et al. (2011), qui soutient l'importance

de connexions cortico-corticales de longue portée dans un système concret qui sous-tend la conscience. Dans la théorie originale, c'est lorsqu'ils activent les connexions éparses de longue portée de l'espace de travail que les processus automatiques ou « habituels/pavloviens », habituellement ségrégués dans des modules aux dynamiques à « dominance componentielle », deviennent des processus délibérés ou « dirigés vers des buts ». Une particularité de l'espace de travail, c'est qu'il s'active par des interactions massivement non-linéaires qui produisent une « ignition » par l'amplification tardive des signaux sensoriels, une augmentation des oscillations de haute fréquence et une synchronie de phases de longue portée (Dehaene, Changeux et Nacache, 2011). Nous pouvons alors comprendre les dynamiques à « dominance interactive » de la MRH comme étant fondamentales aux processus conscients « dirigés vers des buts ». En effet, l'activation de l'espace de travail global va permettre de contraindre les interactions intermodulaires pour former les assemblages flexibles qui sont essentiels pour la planification « dirigés vers des buts ». C'est parce que les processus « dirigés vers des buts » vont impliquer un contrôle « top-down » qui permet de simuler, dans des séquences hippocampiques générées à l'internes, les conséquences contrefactuelles de l'action en fonction de représentations des buts. Ainsi, nous aurions avantage à employer un pluralisme explicatif intégratif lorsque vient le temps d'aborder les interactions hiérarchiques bidirectionnelles dans le cerveau et accorder une importance égale à la ségrégation locale dans des modules, qui reposent sur des dynamiques à « dominance componentielle » et à l'intégration globale dans des hubs connecteurs « rich-club », qui reposent sur des dynamiques à « dominance interactive ».

RÉFÉRENCES

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4), 245-266.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain.* MIT Press.

Anderson, M. (2015). Beyond componential constitution in the brain: Starburst amacrine cells and enabling constraints. *OpenMinds*.

Anderson, M. L. (2016). Précis of after phrenology: neural reuse and the interactive brain. *Behavioral and Brain Sciences*, *39*.

Anderson, M. L., & Finlay, B. L. (2014). Allocating structure to function: the strong links between neuroplasticity and natural selection. *Frontiers in human neuroscience*, *7*, 918.

Anderson, M. L., Richardson, M. J., & Chemero, A. (2012). *Eroding the boundaries of cognition: implications of embodiment 1*. *Topics in cognitive science*, 4(4), 717-730.

Anggraini, D., Glasauer, S., & Wunderlich, K. (2018). Neural signatures of reinforcement learning correlate with strategy adoption during spatial navigation. *Scientific reports*, *8*(1), 1-14.

Baars, B. J. (2007). The global workspace theory of consciousness. *The Blackwell companion to consciousness*, 236-246.

Badcock, P. B. (2012). Evolutionary systems theory: A unifying meta-theory of psychological science. *Review of General Psychology*, *16*(1), 10-23.

Badcock, P. B., Friston, K. J., Ramstead, M. J., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(6), 1319-1351.

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, *59*(4), 381.

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *Behavioral and brain sciences*, *9*(1), 67-90.

Barbieri, M. (2012). Code biology–A new science of life. *Biosemiotics*, *5*(3), 411-437.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review*, *113*(3), 628.

Barton, R. A., & Harvey, P. H. (2000). Mosaic evolution of brain structure in mammals. *Nature*, *405*(6790), 1055-1058.

Bassett, D. S., Yang, M., Wymbs, N. F., & Grafton, S. T. (2015). Learning-induced autonomy of sensorimotor systems. *Nature neuroscience*, *18*(5), 744-751.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695-711.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press.

Bertolero, M. A., Yeo, B. T., & D'Esposito, M. (2015). The modular and integrative functional architecture of the human brain. *Proceedings of the National Academy of Sciences*, *112*(49), E6798-E6807.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417-2444.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in human neuroscience*, *8*, 599.

Campbell, J. O. (2016). Universal Darwinism as a process of Bayesian inference. *Frontiers in Systems Neuroscience*, *10*, 49.

Caporael, L. R. (2001). Evolutionary psychology: Toward a unifying theory and a hybrid science. *Annual review of psychology*, *52*(1), 607-628.

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS biology*, *3*(7), e245.

Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.

Carruthers, P. (2008). Precis of The architecture of the mind: Massive modularity and the flexibility of thought. *Mind & Language*, *23*(3), 257-262.

Charvet, C. J., & Finlay, B. L. (2012). Embracing covariation in brain evolution: large brains, extended development, and flexible primate social systems. *Progress in brain research*, *195*, 71-87.

Cherniak, C., Mokhtarzada, Z., Rodriguez-Esteban, R., & Changizi, K. (2004). Global optimization of cerebral cortex layout. *Proceedings of the National Academy of Sciences*, *101*(4), 1081-1086.

Chomsky, N. (1975). *Reflections on language*. New York: Random House.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585-1599.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Clark, A., & Toribio, J. (1994). *Doing without* representing?. *Synthese*, *101*(3), 401-431.

Cohen, J. R., & D'Esposito, M. (2016). The segregation and integration of distinct brain networks and their relationship to cognition. *Journal of Neuroscience*, *36*(48), 12083-12094.

Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*(2), 89-97.

Constant, A., Clark, A., & Friston, K. J. (2021). Representation wars: Enacting an armistice through active inference. *Frontiers in Psychology*, *11*, 3798.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. *Mapping the mind: Domain specificity in cognition and culture*, 853116.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Cuffari, E. C., Di Paolo, E., & De Jaegher, H. (2015). From participatory sense-making to language: there and back again. *Phenomenology and the Cognitive Sciences*, *14*(4), 1089-1125.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *Journal of Mathematical Psychology*, *99*, 102447.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12), 1704-1711.

Dehaene, S., Changeux, J. P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications. *Characterizing consciousness: From cognition to the clinic?*, 55-84.

Dreyfus, H. L. (2002). *A phenomenology of skill acquisition as the basis for a Merleau-Pontian nonrepresentational cognitive science*. https://philarchive.org/archive/DREAPO

Eliasmith, C. (1997). Computation and dynamical models of mind. *Minds and Machines*, *7*(4), 531-541.

Everitt, B. J., & Robbins, T. W. (2013). From the ventral to the dorsal striatum: devolving views of their roles in drug addiction. *Neuroscience & Biobehavioral Reviews*, *37*(9), 1946-1954.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, *4*, 215.

Finlay, B. L., & Darlington, R. B. (1995). Linked regularities in the development and evolution of mammalian brains. *Science*, *268*(5217), 1578-1584.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Fodor, J. A. (1990). *A theory of content and other essays*. The MIT Press.

Fodor, J. (2001). *Evolution and the human mind: Modularity, language and meta-cognition*. Cambridge University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K. (2018). Am I self-conscious?(Or does self-organization entail self-consciousness?). *Frontiers in psychology*, *9*, 579.

Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends in cognitive sciences*, *20*(7), 500-511.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862-879.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*, *1*(4), 381-414.

Friston, K. J., & Price, C. J. (2011). Modules and brain mapping. *Cognitive neuropsychology*, *28*(3-4), 241-250.

Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction?. *Journal of consciousness studies*, *8*(5-6), 83-108.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. *The shared mind: Perspectives on intersubjectivity*, *12*, 17-38.

Gallistel, C. R. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. *The cognitive neurosciences*, *2*, 1179-1191.

Gibson, J. J. (1977). The concept of affordances. Perceiving, acting, and knowing. Dans R. E. Shaw, J. Bransford (dir.), *Perceiving, acting, and knowing: toward an ecological psychology*. Lawrence Erlbaum Associates.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & philosophy*, *32*(3), 337-355.

Hager, R., Lu, L., Rosen, G. D., & Williams, R. W. (2012). Genetic architecture supports mosaic brain evolution and independent brain–body size regulation. *Nature communications*, *3*(1), 1-5.

Hesse, J., & Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems. *Frontiers in systems neuroscience*, *8*, 166.

Hilgetag, C. C., & Hütt, M. T. (2014). *Hierarchical modular brain connectivity is a stretch for criticality*. Trends in cognitive sciences, *18*(3), 114-115.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2017). How to entrain your evil demon. Dans T. Metzinger & W. Wiese (dir.), *Philosophy and Predictive Processing.* MIND Group.

Huneman, P. (2006). Naturalising purpose: from comparative anatomy to the 'adventure of reason'. *Studies in history and philosophy of biological and biomedical sciences, 37*(4), 649–674.

Huneman, P. (2018). Outlines of a theory of structural explanations. *Philosophical Studies*, *175*(3), 665-702.

Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. *Royal Institute of Philosophy Supplements*, *60*, 43-68.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. MIT press.

Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia*, *43*(3), 521-536.

James, W. (1890). The perception of reality. *Principles of psychology*, *2*, 283-324.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Kant, I. (1987). *Critique of judgment*. Hackett Publishing.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA.

Kello, C. T., & Van Orden, G. C. (2009). Soft-assembly of sensorimotor function. *Nonlinear dynamics, psychology, and life sciences*, *13*(1), 57.

Kelso, J. A. S. (2009). Coordination Dynamics. Dans: *Encyclopedia of Complexity and System Science* (dir.), R. A. Meyers.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387-2415.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, *15*(138), 20170792.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264-281.

Kiverstein, J., & Rietveld, E. (2015). The primacy of skilled intentionality: on Hutto & Satne's the natural origins of content. *Philosophia*, *43*(3), 701-721.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712-719.

Laland, K. N., Odling-Smee, J., & Feldman, M. W. (2000). Niche construction, biological evolution, and cultural change. *Behavioral and brain sciences*, *23*(1), 131-146.

Lewis, M. D. (2000). The promise of dynamic systems approaches for an integrated account of human development. *Child development*, *71*(1), 36-43.

Lundie, M. (2019). Systemic functional adaptedness and domain-general cognition: broadening the scope of evolutionary psychology. *Biology & Philosophy*, *34*(1), 8.

Maisto, D., Friston, K., & Pezzulo, G. (2019). Caching mechanisms for habit formation in active inference. *Neurocomputing*, *359*, 298-314.

Mameli, M., & Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1563), 436-443.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.

Marr, D. (1976). Early processing of visual information*. Philosophical Transactions of the Royal Society of London*. B, Biological Sciences, *275*(942), 483-519.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press.

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, *4*, 200.

Milkowski, M. (2008). When weak modularity is robust enough?*. Análisis filosófico*, *28*(1), 77-89.

Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural brain research*, *6*(1), 57-77.

Mitchell, S. D. (1992). On pluralism and competition in evolutionary explanations. *American Zoologist*, *32*(1), 135-144.

Montgomery, S. H., Mundy, N. I., & Barton, R. A. (2016). Brain evolution and development: adaptation, allometry and constraint. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1838), 20160433.

Moreno, A., & Suárez, J. (2020). Plurality of explanatory strategies in biology: Mechanisms and Network*s. Methodological prospects for scientific research*. PP. 141-165.

Medina, M. L. (2010). Two "EvoDevos". *Biological Theory*, *5*(1), 7-11.

Noë, A., & Noë, A. (2004). *Action in perception*. MIT press.

O'regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.

Oyama, S., Griffiths, P. E., & Gray, R. D. (2001*). Cycles of contingency: Developmental systems and evolution*. Mit Press.

Park, H. J., & Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science*, *342*(6158).

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface*, *14*(136), 20170376.

Parr, T., Sajid, N., & Friston, K. J. (2020). Modules or mean-fields?. *Entropy*, *22*(5), 552.

Pearl, J. (2014). Probabilistic reasoning in intelligent systems: networks of plausible inference. *Elsevier*.

Perrinet, L. U., Adams, R. A., & Friston, K. J. (2014). Active inference, eye movements and oculomotor delays. *Biological cybernetics*, *108*(6), 777-801.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L., & Friston, K. (2016). Active inference, epistemic value, and vicarious trial and error. *Learning & Memory*, *23*(7), 322-338.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, *134*, 17-35.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, *22*(4), 294-306.

Pezzulo, G., van der Meer, M. A., Lansink, C. S., & Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in cognitive sciences*, *18*(12), 647-657.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of biological physics*, *37*(1), 1-38.

Pinker, S. (2003). *The blank slate: The modern denial of human nature*. Penguin.

Raja, V., & Anderson, M. L. (2021). Behavior considered as an enabling constraint. *Neural Mechanisms*. ¨PP. 209-232.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2021). Multiscale integration: beyond internalism and externalism. *Synthese*, *198*(1), 41-70.

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225-239.

Rathkopf, C. (2018). *Network representation and complex systems*. *Synthese*, *195*(1), 55-78.

Richardson, M. J., & Chemero, A. (2014). Complex dynamical systems and embodiment. Dans: L. Shapiro, *The Routledge handbook of embodied cognition*. Routledge.

Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind*, *117*(468), 973-1001.

Safron, A. (2020). An Integrated World Modeling Theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; Toward solving the hard problem and characterizing agentic causation. *Frontiers in artificial intelligence*, *3*, 30.

Schiavio, A., & Kimmel, M. (2021). *The ecological dynamics of musical creativity and skill acquisition*. Dans: A. Scarinzi (dir.). *Meaningful Relations: The Enactivist Making of Experiential Worlds*. Academia-Verlag.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in psychology*, *7*, 1792.

Simon, H. A. (1991). The architecture of complexity. Dans: *Facets of systems science* (pp. 457-476). Springer.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(1), 1-23.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. *Mapping the mind: Domain specificity in cognition and culture*, 39-67.

Sporns, O. (2010). *Networks of the Brain*. MIT press.

Sterelny, K. (2003). *Thought in a Hostile World.* Blackwells.

Sternberg, S. (2011). Modular processes in mind and brain. *Cognitive neuropsychology*, *28*(3-4), 156-208.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410-433.

Tooby, J., & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. Dans: D. M. Buss (dir.), *The Handbook of Evolutionary Psychology, Second edition*, *Volume 1: Foundations*. John Wiley & Sons.

Uexküll, J. V., & Schiller, C. H. (1957). *Instinctive behavior*. New York, International Universities Press.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of experimental psychology: General*, *132*(3), 331.

Wilson R. A., Foglia L., Shapiro L. & Spaulding S. (2021). *Embodied Cognition*. Dans E. N. Zlata (dir.) The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2021/entries/embodied-cognition/

Whiting, B. A., & Barton, R. A. (2003). The evolution of the cortico-cerebellar complex in primates: anatomical connections predict patterns of correlated evolution. *Journal of human evolution*, *44*(1), 3-10.

Witherington, D. C., & Lickliter, R. (2016). Integrating development and evolution in psychological science: Evolutionary developmental psychology, developmental systems, and explanatory pluralism. *Human Development*, *59*(4), 200-234.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880-1882.