UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LA MODULARITÉ DES ADAPTATIONS PSYCHOLOGIQUES HUMAINES : LE CAS DE LA THÉORIE DE L'ESPRIT

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN PHILOSOPHIE

PAR

BERNARD CORDEAU

JANVIER 2015

REMERCIEMENTS

J'aimerais remercier Luc Faucher pour sa présence encourageante qui, au cours des dernières années, fut indispensable à la complétion de ce mémoire. J'aimerais aussi remercier Pierre Poirier pour ses conseils judicieux et sa généreuse disponibilité. Ce mémoire est en grande partie redevable des nombreuses conversations que j'ai pues entretenir avec Pierre et Luc depuis mon arrivée à l'Université du Québec à Montréal.

Finalement, je suis reconnaissant des commentaires et suggestions de Serge Robert sur certaines sections de ce mémoire.

À Julianne, pour tout

TABLES DES MATIÈRES

RÉSUMÉ

Dans le but de supporter la plausibilité de la version adaptationniste de l'hypothèse de la modularité massive de l'esprit (aHMM), nous démontrerons qu'une capacité psychologique de haut-niveau, la capacité d'attribution d'états mentaux épistémiques (éTdE), est sous-tendue par un module darwinien. Avant tout, nous caractériserons la notion de module impliquée par l'aHMM (i.e. la modularité darwinienne) en la distinguant des principaux usages de la modularité en sciences cognitives (i.e. la conception modulaire de Marr, la modularité chomskyenne et la modularité fodorienne). Ensuite, nous démontrerons que l'éTdE est bel et bien sous-tendue par une structure psychologique qui satisfait les critères de la modularité darwinienne. Pour ce faire, nous établirons, en premier lieu, l'existence d'une région corticale fonctionnellement spécialisée pour l'éTdE et, en second lieu, que cette région corticale particulière est le produit de la sélection naturelle. De manière générale, cette démonstration milite en faveur de l'idée que la postulation de modules darwiniens peut contribuer heuristiquement à la décomposition de la cognition humaine.

MOTS-CLÉS : cognition sociale, évolution de la cognition, évolution humaine, modularité darwinienne, psychologie évolutionniste.

INTRODUCTION

Pour la seconde édition (1845) de son journal de voyage (communément appelé *The Voyage of the Beagle*), Charles Darwin effectua une révision importante de ses notes. Il y ajouta, entre autres, sa désormais célèbre hypothèse expliquant le lien entre la distribution géographique et la morphologie distinctive des pinsons des Galápagos : « Seeing this gradation and diversity of structure in one small, intimately related group of birds, one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends. » (Darwin, 1845, p. 380). Déjà, cette conjecture dénotait de la reconnaissance de deux leçons fondamentales de la théorie darwinienne d'évolution par sélection naturelle. D'un côté, la production de nouvelles espèces par modification graduelle des traits, expliquant comment la taille et la forme du bec des différentes espèces de pinsons étaient adaptées aux pressions sélectives de leurs niches écologiques respectives. De l'autre, la conservation des traits par descendance à partir d'un ancêtre commun, expliquant la ressemblance intime, hormis le bec, des différentes espèces de pinsons. De manière générale, dans ce mémoire, nous tenterons de mettre en évidence l'importance de ces deux leçons pour la décomposition de la cognition humaine « à ses joints ».

Il est généralement admis que l'encéphalisation humaine est le produit de la sélection naturelle. Il s'agirait d'une adaptation *anatomique* humaine. La question de l'existence d'adaptations *psychologiques* humaines est, quant à elle, plus contentieuse. En effet, le fait que la cognition humaine soit le résultat de l'évolution biologique n'implique pas que les propriétés des différentes structures psychologiques la composant soient des produits de la sélection naturelle. Par exemple, un gros cerveau pourrait avoir des capacités différentes d'un petit cerveau, sans que ces capacités aient été

spécifiquement sélectionnées. De plus, même s'il s'avérait exact que certaines structures de la cognition humaine furent spécifiquement sélectionnées, cela ne permettrait pas de déterminer *a priori* le nombre, l'organisation ou la nature de ces structures. Afin d'établir la plausibilité de la version adaptationniste de l'hypothèse de la modularité massive de l'esprit (aHMM) développée par certains psychologues évolutionnistes – thèse selon laquelle plusieurs structures de la cognition humaine de haut-niveau sont des adaptations spécialisées dans la résolution d'un problème évolutif spécifique – nous proposons d'établir qu'une sous-capacité de haut-niveau de la cognition sociale humaine est effectivement sous-tendue par un module darwinien (i.e une structure psychologique, cognitive ou neurologique, fonctionnellement spécialisée qui fut « façonnée » (*designed*) par la sélection naturelle). Notre objectif est d'établir l'idée que la postulation de modules darwiniens peut contribuer heuristiquement à la décomposition de la cognition humaine.

Pour ce faire, nous proposons, dans le premier chapitre, une analyse philosophique des rôles et des postulats associés aux principaux usages de la modularité en sciences cognitives (i.e. la conception modulaire de Marr, la modularité chomskyenne et la modularité fodorienne). Cette analyse nous permettra à la fois de constater la disparité des usages de la modularité en sciences cognitives et de les distinguer du concept de modularité auquel s'engagent les partisans de l'aHMM (i.e. le concept de modularité darwinienne). Par la suite, nous caractériserons cette modularité darwinienne en établissant le rôle théorique ainsi que les postulats modulaire qui lui sont associés. En particulier, nous expliciterons le postulat primaire propre à cet usage de la modularité : la spécialisation fonctionnelle évoluée. L'analyse détaillée des implications de la spécialisation fonctionnelle évoluée au niveau cognitif et neurologique, nous permettra d'établir la nature distincte de la modularité darwinienne. À la lumière de cette caractérisation de la modularité darwinienne, nous nous opposerons à l'idée que les modules darwiniens, en abandonnant les propriétés associées aux modules fodoriens, se réduiraient alors trivialement aux composantes

postulées par la décomposition fonctionnelle.

Dans le second chapitre, cette caractérisation, au niveau cognitif et neurologique, de la spécialisation fonctionnelle évoluée, nous permettra de soutenir qu'il est plausible que la capacité d'attribution d'états mentaux épistémiques (éTdE) soit sous-tendue par un module darwinien. Pour démontrer ce dernier point, nous examinerons, grâce à une revue de la littérature sur la théorie de l'esprit (TdE), en premier lieu, si un ensemble de données convergentes supportent l'existence d'une structure psychologique spécialisée pour l'éTdE et, en second lieu, s'il est plausible que cette structure psychologique soit un produit de la sélection naturelle. Nous sommes conscients que même si s'avérait exact que l'éTdE soit sous-tendue par un module darwinien, cela ne confirmerait pas définitivement la véracité de l'aHMM. Nous croyons toutefois que cette éventualité militerait fortement en faveur de, non seulement, la plausibilité de l'aHMM, mais aussi la valeur heuristique de la modularité darwinienne pour la décomposition de la cognition humaine.

CHAPITRE I

LES MODULES DARWINIENS

> « But, clearly, what one ought to say (....) about whether the mind is massively modular, depends on what one takes a module to be. »
>
> (Fodor, 2000, p. 56)

1.1. Les usages de la modularité en sciences cognitives

Les termes « modularité », « module » et « modulaire » sont utilisés dans plusieurs disciplines académiques (notamment la biologie, l'ingénierie et les mathématiques) pour désigner divers types d'objets et de propriétés (pour un aperçu de l'étendue de cette diversité, voir l'anthologie de Callebaut et Rasskin-Gutman, 2005). Dans ce mémoire, nous nous intéressons exclusivement aux usages présents en sciences cognitives. De manière générale, en sciences cognitives, les *modules* désignent un type particulier de structures (anatomiques ou fonctionnelles) isolables et relativement indépendantes de la cognition humaine, et les *architectures modulaires* désignent toute organisation constituée, du moins en partie, de ces structures particulières. Il est toutefois difficile de fournir une caractérisation inclusive des modules de l'esprit/cerveau, car toute tentative en ce sens est rapidement confrontée à l'absence de consensus à propos de l'usage et de l'application de la modularité en sciences cognitives. Comme le constate Seok : « the most serious challenge to modularity comes from the complexity of modularity. (…) there are so many features, so many dimensions, and so many modules proposed and discussed. But there seems to be no theoretical unity in the use and application of modularity in cognitive science. » (2006, p. 366). En effet, suite à l'influente contribution de Fodor

(1983), mais aussi, dans une moindre mesure, grâce aux contributions de David Marr (1976, 1982), de Noam Chomsky (1980) et de Tim Shallice (1988), le terme de « module » fut adopté dans plusieurs disciplines des sciences cognitives (e.g., la psychologie cognitive, la linguistique, les neurosciences cognitives et la neuropsychologie cognitive). Parmi ces différentes disciplines, il est possible de recenser plusieurs usages distincts de la notion de module. Cette disparité d'usage permet de répertorier divers types de structures et de propriétés caractérisant les modules en sciences cognitives. Par exemple, selon Coltheart (1999), les modules peuvent être définis comme des systèmes cognitifs répondant exclusivement à une classe particulière de stimuli. Selon Sternberg (2001, 2011), les modules sont des sous-processus cognitifs ou neuronaux séparément modifiables (i.e. un stade distinct de traitement de l'information identifié à partir d'une expérimentation démontrant l'influence sélective de ce sous-processus par une certaine variable)[1]. Selon Calabretta et Parisi (2005), les modules connexionnistes sont des parties anatomiquement séparables et/ou fonctionnellement spécialisées d'un réseau de neurone. Selon Op de Beeck, Haushofer et Kanwisher (2008) ainsi que Kanwisher (2010), les modules sont des parties de régions cérébrales discrètes contribuant de manière sélective aux traitement d'une classe spécifique de stimuli. Selon Meunier, Lambiotte et Bullmore (2010), les modules topologiques sont des réseaux de connexions et d'activités corticales, décrits mathématiquement, caractérisés par différentes propriétés structurelles et fonctionnelles typiques des hiérarchies quasi-décomposables. On dénombre d'ailleurs plusieurs tentatives de distinction et d'explicitation de certains des usages de la modularité ayant cours en sciences cognitives (e.g., Bechtel, 2003; Bergeron, 2007, 2008; Faucher et Tappolet, 2006; Faucher et Poirier, 2009; Mahon et Cantlon, 2011; Robbins, 2010; Segal, 1996; Samuels, 1998, 2000, 2006; Seok, 2006). Exception faite de l'analyse de Bergeron (2007, 2008), qui identifie deux dimensions ontologiques communes (i.e. la dimension anatomique et la dimension fonctionnelle)

[1] Voir Coltheart (2011) pour une comparaison entre la notion sternbergienne de module et la notion fodorienne de module.

aux principales formes de modularité utilisées en sciences cognitives, une revue de cette littérature indique que les divers usages de la modularité impliquent une variété de postulats modulaires distincts portant sur un large éventail de dimensions ontologiques (e.g. le type de structure biologique, le type de fonction cognitive, le type de traitement de l'information, le type d'ontogénie, le type de phylogénie ou encore le type de correspondance structure/fonction). En somme, nous n'identifions pas de caractérisation générale ou unifiée de la modularité en sciences cognitives. Au contraire, nous constatons une multiplication grandissante d'usages disparates.

Dans ce chapitre, nous nous limiterons à souligner, de manière non exhaustive, certaines similarité et différences des principaux usages de la modularité en sciences cognitives. Tout d'abord, nous présenterons les usages que Fodor (1983), Chomsky (1980) et Marr (1976, 1982) font respectivement de la notion de module. Lors de cette présentation, nous identifions les postulats associés à chacun de ces usages, et soulignons le rôle théorique joué par chacun de ces usages dans leur programme de recherche respectif. Cette analyse nous permettra, par la suite, d'expliciter le postulat primaire propre à la modularité darwinienne : la spécialisation fonctionnelle évoluée.

1.1.2. Les postulats et les inférences modulaires

Avant d'examiner les détails de ces trois usages, il est important de se munir de concepts permettant de déterminer leurs points communs et leurs différences. En premier lieu, nous distinguons deux éléments théoriques en fonction desquels un usage particulier de la modularité peut varier. Il ne faut pas confondre les *postulats modulaires* caractérisant un usage particulier de la modularité, avec les *inférences modulaires* recrutées par cet usage. Afin de clarifier les rôles respectifs des postulats modulaires et des inférences modulaires, nous nous attarderons à l'organologie développée par le neurologue Franz Joseph Gall.

Le cadre théorique de l'organologie de Gall est considéré comme un précurseur des

usages contemporains de la modularité en sciences cognitives (Bergeron, 2008; Fodor, 1983; Zawidski et Bechtel, 2004). De manière générale, Gall proposait d'inférer le degré de développement et la localisation des facultés mentales à partir des caractéristiques morphologiques du crâne et du visage. Par exemple, à partir de l'observation que les individus ayant une mémoire développée avaient les yeux exorbités, Gall (1822-25) inféra que les différentes facultés de mémorisation devaient siéger dans les lobes frontaux. Selon Bergeron (2008), il est important de distinguer le postulat modulaire de Gall des inférences modulaires propres à sa méthodologie phrénologique. D'une part, le postulat modulaire de Gall est que l'esprit humain est constitué de nombreuses facultés mentales indépendantes les unes des autres, chacune étant située dans une région particulière du cerveau. D'autre part, la méthodologie de Gall consiste à inférer la localisation et le degré de développement individuel de ces différentes facultés mentales à partir du relief des boîtes crâniennes. Cette distinction permet d'identifier les raisons pour lesquelles les conclusions phrénologiques de Gall sont erronées, sans avoir à discréditer son postulat modulaire. En effet, il est admis que la méthodologie de Gall (i.e. la phrénologie) recrutait des stratégies inférentielles (i.e. les inférences phrénologiques) qui, à leur tour, reposaient sur des postulats ontologiques qui se sont révélés faux. Par exemple, Gall assumait erronément que la morphologie crânio-faciale correspondait suffisamment à la forme du cerveau pour que l'on puisse inférer le volume des circonvolutions cérébrales à partir de la forme du crâne, ou encore que le degré de développement des facultés mentales était corrélé avec le volume des circonvolutions (*gyrus*) cérébrales responsables de ces facultés.

De nos jours, l'élucidation de la nature spécifique de la modularité du système cognitif humain s'effectue à l'aide de stratégies inférentielles opérant, non pas à partir de mesures du relief crânien, mais à partir, entre autres, de mesures des performances comportementales humaines à des tâches cognitives variées. Ainsi, selon Bergeron (2007, 2008), les *postulats modulaires* sont les hypothèses spécifiant les propriétés générales des modules ou de leur organisation et les *inférences modulaires* sont les

raisonnements et les méthodes déployés afin de découvrir ou de confirmer la nature spécifique des modules ou de leur organisation. Il est possible d'identifier différentes inférences modulaires associées à différentes disciplines ou programmes de recherche en sciences cognitives. Certaines tentent d'inférer les propriétés des structures (anatomiques ou fonctionnelles) particulières sous-jacentes aux fonctions psychologiques spécifiques, tandis que d'autres tentent d'inférer les propriétés des fonctions spécifiques réalisées par ces structures particulières.

Pour ajouter à cette distinction, nous croyons qu'il est important de ne pas confondre les *postulats modulaires primaires*, qui spécifient les propriétés caractérisant le type particulier de structures psychologiques, avec les *postulats modulaires secondaires*, qui spécifient les propriétés étant typiquement ou possiblement possédées par ces structures particulières. En effet, une revue de la littérature indique que l'absence de consensus à propos de l'usage de la modularité en sciences cognitives est alimentée, entre autres, par la présence de plusieurs interventions partisanes stipulant soit des types distincts de structures psychologiques (i.e. désaccord sur les postulats modulaires primaires), soit des ensembles distincts de propriétés pouvant, typiquement ou possiblement, être exhibées par les modules (i.e. désaccord sur les postulats modulaires secondaires). Dans les sections suivantes, en explicitant les postulats modulaires de Marr (1976, 1982), de Chomsky (1980) et de Fodor (1983) nous verrons d'ailleurs que les disparités entre les postulats modulaires de Marr et ceux de Fodor concernent principalement les postulats modulaires secondaires, alors que les disparités entre les postulats modulaires de Chomsky et de Fodor concernent essentiellement les postulats modulaires primaires.

1.1.3. Le principe de conception modulaire de Marr

L'un des premiers usages de la notion de modularité en sciences cognitives se retrouve dans le contexte théorique du modèle computationnel des stades primaires de

la vision humaine Marr (1976). Marr, partageant les leçons qu'il tira de ses travaux, propose quatre principes devant guider la modélisation et l'implémentation des processus cognitifs humains. Un de ces principes, le principe de conception modulaire (*principle of modular design*), préconise qu'il est préférable pour modéliser un processus cognitif complexe (e.g. le processus responsable de la vision humaine) de séparer ce dernier et de l'implémenter en une collection de sous-systèmes relativement indépendants les uns des autres (voir également Marr, 1982, p. 102). Autrement dit, Marr considère fructueux d'implémenter chaque partie isolable (i.e. les modules) d'un processus de traitement de l'information dans une structure particulière relativement indépendante.

On peut identifier trois postulats modulaires secondaires caractérisant les modules de Marr. Premièrement, la spécificité fonctionnelle de ces structures à un sous-processus particulier. Deuxièmement, « l'isolabilité » (*isolability*) des processus de traitement de l'information. Troisièmement, la relative indépendance des structures sous-jacentes à ces processus isolables. Notons que cette indépendance relative n'empêche pas que se déroulent de faibles interactions entre les différentes structures qui participent au processus général. En effet, pour Marr, ce qui importe est que ces interactions soient suffisamment faibles pour qu'il soit possible de les étudier de manière indépendante (i.e. qu'il soit possible de spécifier leurs propriétés computationnelles, algorithmiques et implémentationnelles).

Marr présente deux arguments pour, à la fois, motiver l'application du principe de conception modulaire à la vision humaine, et justifier la modularité du système cognitif. Premièrement, s'inspirant de l'argument de Simon (1962), selon lequel la quasi-décomposabilité (*near-decomposability*) serait une propriété importante des systèmes biologiques complexes, Marr (1976) soutient :

> 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).

En d'autres termes, si la modularité n'était pas une propriété de l'organisation neurologique et cognitive, alors l'évolution biologique de la cognition humaine serait, non pas nécessairement impossible, mais du moins improbable (pour une explicitation de cet argument, voir Carruthers, 2006, chap. 1).

Deuxièmement, Marr (1982) remarque que si l'on isole expérimentalement un sous-processus visuel des autres sous-processus participant à la vision et que l'on observe que le fonctionnement de ce sous-processus n'est pas perturbé par cette isolation artificielle, alors il est possible d'inférer que le sous-processus en question est effectivement réalisé par une structure fonctionnellement indépendante. En d'autres termes, la nature modulaire du système nerveux visuel serait démontrée par le constat expérimental que certains processus de la vision primaire ne nécessitent pas d'interactions complexes avec d'autres processus visuels pour accomplir leur fonction. Ainsi, la modularité des modèles computationnels ne serait pas une propriété artificielle, mais correspondrait, à un premier niveau d'approximation, à la nature modulaire du systèmes nerveux.

Une des contributions les plus importantes de Marr aux sciences cognitives réside dans son hypothèse méthodologique selon laquelle un processus de traitement de l'information s'analyse sur trois niveaux complémentaires : le niveau computationnel (spécifiant le problème computationnel que le processus doit résoudre, sa tâche), le niveau algorithmique (spécifiant les représentations et les manipulations sur ces dernières) et le niveau implémentationnel (spécifiant comment et dans quelles structures physiques les opérations algorithmiques sont réalisées). Sans entrer dans les détails de cette analyse à trois niveaux, il est important de souligner que la modularité est fondamentale à cette approche. En effet, selon Marr (1982, p. 356), sans l'existence de sous-processus isolables, nous avons peu de chance de i) découvrir les contraintes computationnelles permettant de développer une théorie

computationnelle spécifiant la fonction d'un processus cognitif (niveau computationnel), ii) développer l'algorithme spécifiant les opérations et les représentations par lesquelles s'effectuent cette fonction cognitive (niveau algorithmique) et iii) reproduire ou localiser les structures implémentant cette fonction cognitive. Autrement dit, dans le cadre théorique de Marr, un des rôles de la modularité du système cognitif est de légitimer l'application de l'approche computationnelle de la vision.

Ce rôle de légitimation de l'approche computationnelle est un rôle récurrent de la modularité en sciences cognitives. Comme nous le verrons (sous-section 1.1.5), il est aussi présent chez Fodor (1983, 2000), car il oriente la majorité du débat sur la possibilité de l'application de l'approche computationnelle aux capacités cognitives centrales. Dans les sous-sections 1.1.4 et 1.1.5, nous nous limiterons à examiner le rôle distinct que Chomsky et Fodor accordent à leur forme respective de modularité, ainsi qu'à distinguer les postulats modulaires qui y sont associés.

1.1.4. Modularité chomskyenne

Comme son nom l'indique, la modularité chomskyenne tire son origine des travaux de Chomsky en linguistique. Précisément, la notion de module chomskyen, développée par Samuels (1998; voir également Fodor, 2000), caractérise un type de structure cognitive qui eut une grande influence en psychologie cognitive, particulièrement en psychologie développementale (Hirschfeld et Gelman, 1994), et dont l'exemple paradigmatique est la notion de « grammaire générative ». Dans le cadre de sa théorie du langage, Chomsky inférait l'existence d'une structure mentale – un organe mental - qui rend compte de la *compétence* linguistique des locuteurs d'un langage. Cette structure mentale est un système de règles qu'il nomme « grammaire générative » (Chomsky, 1980). En plus de participer à la compréhension et à la production linguistique, cette grammaire informe les jugements linguistiques qu'un locuteur peut

effectuer. Ainsi, selon Chomsky, les structures mentales peuplant l'esprit sont des ensembles de connaissances ou de contenus propositionnels spécifiques à un domaine (e.g. le langage) (Fodor, 1983).

Outre le postulat modulaire primaire selon lequel les modules chomskyens sont des ensembles de connaissances ou de contenus propositionnels, il est possible d'identifier trois postulats modulaires secondaires caractérisant la modularité chomskyenne. Premièrement, chacun de ces ensembles de connaissances ou de contenus propositionnels serait spécifique à un domaine de connaissance distinct. Deuxièmement, ces ensembles de connaissances ne seraient pas accessibles à la conscience. Par exemple, nous n'aurions pas accès aux contenus représentationnels de notre « grammaire générative ». Troisièmement, ils seraient innés, ou du moins, ils découleraient d'un développement endogène. Notons qu'il pourrait être restrictif d'identifier d'autres postulats modulaires secondaires pour caractériser les modules chomskyens, car, selon Chomsky (1980, p. 27), il est probable que les principes d'un domaine de connaissance spécifique (e.g. langage), déterminant les propriétés du module chomskyen qui le sous-tend (e.g. la « grammaire générative »), ne s'apparentent pas aux principes des autres domaines de connaissances susceptibles d'être sous-tendus par un module chomskyen. Nous reviendrons sur une notion similaire lorsque nous aborderons, dans la sous-section 1.2.2, l'hétérogénéité des propriétés des différents modules darwiniens. En effet, nous verrons qu'il est plausible que la découverte des propriétés d'une structure spécialisée particulière de la cognition humaine ne fournisse pas d'information sur les propriétés des autres structures spécialisées peuplant l'esprit humain.

1.1.5. Modularité fodorienne

Dans *Modularity of Mind*, Fodor (1983) propose, à partir d'une revue de l'état des sciences cognitives du langage, de l'audition et de la vision, un modèle général de

l'architecture cognitive humaine, ainsi qu'une caractérisation détaillée de la notion de module. Il amorce cette caractérisation en distinguant les structures mentales qui l'intéressent – les modules fodoriens – des modules chomskyens. Selon Fodor (1983, p. 10), un module fodorien constitue une structure mentale individuée fonctionnellement, c'est-à-dire « by reference to its typical effects ». En comparaison, les modules chomskyens sont individués « by reference to their typical propositional contents ». Précisément, les modules fodoriens sont, comme les modules de Marr, des mécanismes de traitement de l'information. Autrement dit, Fodor et Marr, contrairement à Chomsky, adoptent le même postulat modulaire primaire.

Ce postulat modulaire primaire permet à Fodor de préciser le type de traitement de l'information que les modules fodoriens effectuent. Pour ce faire, Fodor s'inspire de la décomposition fonctionnelle de Gall[2] pour distinguer les traitements de l'information s'effectuant sur plusieurs classes, ou une classe générale, d'information ou de stimulus (e.g. la perception, le jugement, la mémorisation, l'attention, l'imagination), des traitements de l'information s'effectuant sur une classe spécifique d'information ou de stimulus (e.g. perception des visages, reconnaissance de la parole). Selon lui, la psychologie s'intéresse traditionnellement à la décomposition des capacités mentales en différentes capacités générales. En revanche, les modules fodoriens seraient caractérisés par le fait qu'ils traitent typiquement des classes particulières d'information ou de stimulus. Par exemple, dans le cas hypothétique d'une décomposition fonctionnelle de la capacité de mémorisation humaine en différents modules fodoriens, chacun de ces modules serait distinct en fonction du domaine particulier d'information qu'il traite (e.g. les lieux, les visages), et non en fonction du type de traitement qu'il effectue sur ces informations (e.g. mémorisation à court terme, mémorisation à long terme). Il est d'ailleurs plausible que les modules fodoriens soient typiquement dédiés au traitement de modules chomskyens (Fodor,

[2] Pour une discussion de la récupération de la décomposition fonctionnelle de Gall par Fodor, voir Zawidski et Bechtel, 2004.

2000, p. 57; Coltheart 1999, p. 118)[3].

Fodor (1983) développe une architecture cognitive constituée de quatre types de composantes fonctionnelles : les transducteurs, les systèmes d'analyse d'entrées, les systèmes centraux responsable de l'intégration de l'information et les systèmes moteurs-exécuteurs. Une particularité fondamentale de cette architecture modulaire réside dans la distinction entre deux types de structures cognitives : les systèmes d'entrées et de sorties et les systèmes centraux. En effet, Fodor considère que les systèmes périphériques (i.e. les systèmes d'entrées et de sorties) sont des structures particulières exhibant typiquement un ensemble de propriétés cooccurentes qui les distinguent des systèmes centraux. Plus précisément, Fodor (1983) identifie explicitement neuf propriétés habituellement cooccurrentes caractérisant plus ou moins[4] ces systèmes périphériques. Suite à l'identification de certaines propriétés computationnelles, architecturales, neurologiques et développementales exhibées par les systèmes périphériques, un des objectifs de Fodor (1983) consiste à convaincre que la cooccurence de ces dernières s'explique par le fait que ces systèmes constituent une espèce naturelle (*natural kind*) : les modules fodoriens. Selon notre cadre d'analyse, ces propriétés habituellement cooccurentes des mécanismes périphériques de traitement de l'information sont les postulats modulaires secondaires de la modularité fodorienne, car elles spécifient les caractéristiques typiquement exhibées par les structures mentales qui intéressent Fodor (i.e. les mécanismes périphériques de traitement de l'information). Nous reviendrons toutefois plus bas sur le statut primordial d'une de ces propriétés : l'encapsulation informationnelle.

1- Spécificité à un domaine informationnel : Fodor (1983, p. 48) soutient que la notion de spécificité à un domaine informationnel qui l'intéresse est l'intuition initiale de Gall selon laquelle des processus psychologiques distincts correspondent à des domaines de stimulus distincts. Il identifie certains

[3] Il demeure possible que les modules chomskyens soient traités par des mécanismes de traitement de l'information qui ne sont pas des modules fodoriens.

[4] La possession de ces propriétés par les modules admet des degrés (Fodor, 1983, p. 37).

indices empiriques, tels que l'activation sélective d'un processus à une classe spécifique d'information, permettant de détecter cette spécificité à un domaine. Dans *The Mind Doesn't Work That Way (TMDW)*, Fodor (2000, p. 58-61) revient sur cette notion de spécificité à un domaine et soutient que cette spécificité à un domaine informationnel, pour éviter d'être triviale, ne peut ni être une propriété exclusivement relative aux informations, ni être une propriété exclusivement relative aux processus. Par conséquent, Fodor (2000, p. 60) considère qu'il s'agit d'une propriété relationnelle s'appliquant à la manière dont l'information et les processus computationnels interagissent. Pour l'instant, nous soulignerons, suivant Barrett (2005), que la notion de spécificité à un domaine de Fodor (1983) est une notion complexe amalgamant plusieurs types de spécificité (e.g. spécificité d'accès, spécificité de traitement, spécificité fonctionnelle). En effet, Fodor (1983) suggère que l'on peut inférer la spécificité à un domaine d'un processus à la fois à partir de l'observation d'une activation sélective d'un processus par une classe d'information spécifique (i.e. spécificité d'accès), ou encore par le fait qu'un processus opère sur un domaine d'information restreint ou excentrique (i.e. spécificité de traitement). À d'autres moments, il exemplifie la spécificité à un domaine en remarquant qu'un processus peut être ajusté (*closely tuned*) pour des propriétés de stimulus spécifique ou encore contenir une théorie élaborée des objets de son domaine informationnel.

2- Obligatoire : Un processus est obligatoire lorsqu'il est automatique et autonome. Ses opérations ne peuvent donc pas être contrôlées ou modifiées par des processus de haut niveau. Ce qui entraîne que ses opérations ne sont modifiables que par une perturbation des transducteurs.

3- Impénétrabilité cognitive : De manière générale, un processus cognitif est impénétrable aux yeux d'un autre processus si ce dernier n'a pleinement accès qu'aux sorties finales d'un processus et non pas à ses représentations

intermédiaires. Fodor conçoit l'impénétrabilité cognitive des modules en rapport aux processus centraux, notamment en rapport aux processus conscients. Ainsi, un processus serait impénétrable s'il utilise « à l'interne » des informations qui sont inaccessibles aux systèmes centraux (ou à la conscience).

4- Rapidité de traitement : Mesurée par le temps qui s'écoule entre la présentation d'une entrée et la production d'une sortie, Fodor (1983, p. 64) suggère que la rapidité de traitement d'un processus découle de l'obligation de ce traitement, car un processus obligatoire ne perdrait pas de temps à déterminer s'il doit, et comment il doit, traiter ses entrées.

5- Encapsulation informationnelle : Pour Fodor (2000, p. 63), l'encapsulation est la propriété centrale de la modularité cognitive. Cette propriété implique qu'un processus n'a accès qu'à un ensemble restreint d'informations d'arrière-plan lors du traitement de ses entrées. De plus, cette propriété doit être architecturalement imposée puisque les restrictions sur la circulation des informations vers les différents mécanismes encapsulés ne doivent pas être le résultat de facteurs de performance (e.g. fatigue, contrainte temporelle ou inattention) ou de toutes autres causes exclusivement psychologiques (e.g. modifications des croyances, des buts ou des connaissances) (voir Samuels, 2005). Fodor associe l'encapsulation informationnelle d'un processus à la notion d'autonomie informationnelle. Un processus autonome informationnellement n'a pas recours aux informations manipulées par d'autres processus. Autrement dit, un processus autonome informationnellement ne nécessite pas d'échange ou de partage d'information. Dans le cas d'un processus encapsulé, ce dernier n'a typiquement accès qu'à sa base de données privée (*proprietary database*) pour accomplir sa fonction cognitive. Ce processus est alors considéré encapsulé informationnellement par rapport à toute information qui ne provient pas de cette base de données,

en particulier par rapport aux informations provenant des processus centraux.

6- Architecture neuronale fixe : Selon Fodor (1983, p. 98-99), l'architecture neuronale fixe est la concomitante implémentationnelle de l'encapsulation informationnelle. Elle permet d'expliquer l'accès différentiel d'un module fodorien aux outputs des autres processus cognitifs (i.e. l'encapsulation relative de ces modules) en se basant sur l'hypothèse que les connections aux autres processus sont « hard-wired » et forment des voies privilégiées d'accès à l'information. Cependant, cet isomorphisme[5] entre l'organisation des structures cognitives et l'organisation des structures neurologiques n'implique pas que chaque module fodorien soit implémenté dans une région corticale discrète.

7- Dysfonctionnement spécifique suite à une lésion : Considérant que les modules fodoriens sont habituellement réalisés par des circuits neuronaux fixes, nous devrions nous attendre à ce que l'observation d'une lésion de ces circuits neuronaux implémentant un module fodorien entraine un dysfonctionnement comportemental spécifique (e.g. phonoagnosie, prosopagnosie). Précisément, Fodor considère que l'observation éventuelle d'un déficit sélectif d'une capacité générale (e.g. attention, mémoire) suite à une lésion corticale militerait contre la nature modulaire du processus cognitif la réalisant.

8- Superficialité des sorties : La superficialité des sorties est liée aux contraintes informationnelles découlant de l'encapsulation des modules fodoriens: plus un mécanisme est encapsulé, plus ses opérations ont accès à une quantité réduite d'informations, et plus la probabilité qu'ils fournissent des sorties superficielles est élevée. Faucher et Tappolet (2006) remarquent que cette notion de sorties superficielles est associée, par certains auteurs (e.g.

[5] « It is, in short, reasonable to expect biases in the distribution of information to mental process to show up as structural biases in neural architecture » (Fodor, 1983, p. 118).

Carruthers, 2006), à la notion de sorties non-conceptuelles.

9- Programme caractéristique de développement ontogénétique : Chaque module fodorien suivrait un programme robuste de maturation spécifique. Autrement dit, le parcours développemental des structures neurologiques réalisant les modules fodoriens suivrait des étapes caractéristiques déterminées par des facteurs endogènes, mais pouvant être influencées par des déclencheurs environnementaux.

Selon Coltheart (1999), Fodor (1983) ne considère pas ces propriétés comme des conditions nécessaires et suffisantes de la modularité. En d'autres termes, ces propriétés ne seraient pas définitionnelles et un module fodorien ne serait pas tenu de posséder l'ensemble de ces propriétés. Il est toutefois clair que Fodor accorde à l'encapsulation informationnelle une place primordiale (Fodor 2000, p. 63). Dans le reste de cette section, nous mettrons en évidence les raisons pour lesquelles l'encapsulation informationnelle est une notion fondamentale pour la modularité fodorienne. Pour ce faire, nous examinerons le rôle que Fodor (1983, 2000) accorde à la modularité fodorienne dans la résolution du problème du cadre (*frame problem*) auquel fait face le computationnalisme classique.

Une thèse importante de Fodor (1983) avance que les progrès concernant l'analyse cognitive de la vision, de l'audition et du langage découlent principalement de la nature modulaire des systèmes sous-jacents à ces capacités psychologiques. Conversement, Fodor (1983, p. 38) suggère que l'absence de progrès dans l'analyse cognitive des capacités centrales s'explique par le fait que ces dernières ne sont probablement pas sous-tendues par des processus modulaires[6]. Pour justifier cette thèse de la non-modularité des systèmes centraux, Fodor soutient que la fixation intelligente des croyances est la capacité qui caractérise le mieux nos systèmes centraux et que cette fixation intelligente des croyances serait accomplie par des

[6] Pour une critique de la thèse fodorienne de la non-modularité des systèmes centraux, voir Shallice (1984). Pour une critique plus récente, voir Carruthers (2006).

inférences rationnelles qui doivent être sensibles à l'ensemble du système de croyances (i.e. aux propriétés globales du système de croyances). Par exemple, selon Fodor, les inférences rationnelles permettant la fixation des croyances doivent pouvoir évaluer la cohérence d'une croyance par rapport aux autres croyances adoptées par l'organisme. Si l'on ajoute à cela le postulat modulaire secondaire selon lequel les modules fodoriens sont informationnellement encapsulés, il s'ensuit alors que les systèmes centraux ne peuvent être « plausibly viewed as modular ». (Fodor, 1983, p. 103). En effet, un mécanisme encapsulé informationnellement ne peut, par définition, effectuer d'inférences sensibles aux propriétés globales du système de croyances. Dans *TMDW,* Fodor (2000) explicite cet argument (i.e. le problème de la globalité) en identifiant les limitations intrinsèques de la théorie computationnelle de l'esprit (TCE) comme la justification de la non-modularité des systèmes centraux[7]. Précisément, constatant la nature globale des capacités abductives humaines (i.e. les capacités d'intégration de l'information des systèmes centraux), Fodor (2000) soutient l'insuffisance de la TCE, et de la modularité, pour expliquer ces capacités abductives. En effet, selon Fodor (2000), les computations de la TCE sont syntactiquement déterminée (i.e. elles sont uniquement sensibles aux propriétés syntactiques locales d'une représentation) et ne peuvent pas rendre compte de la nature globale des capacités abductives centrales sans faire face au problème du cadre. Le problème du cadre est considéré, notamment par Fodor (1983, 2000), comme la source principale de l'intractabilité[8] des processus computationnels. De manière sommaire, le problème du cadre réfère aux difficultés rencontrées par un système computationnel ayant à déterminer quelles représentations ou informations sont pertinentes à l'accomplissement de sa fonction. Selon Fodor (2000), un processus encapsulé informationnellement évite le problème du cadre puisqu'il obtient, comme

[7] Pour des critiques de cette formulation fodorienne du « problème de la globalité », voir Ludwig et Schneider (2006) ainsi que Samuels (2005, 2010).

[8] La notion de tractabilité computationnelle possède une variété d'usage disparate. Dans ce mémoire, nous adoptons la caractérisation générale de Samuels (2010) selon laquelle la formalisation computationnelle tractable d'une opération mentale ne doit pas nécessiter plus de temps ou de ressources que ce qu'il est raisonnable qu'un humain possède.

nous l'avons mentionné plus haut, de manière architecturalement pré-spécifiée, les informations qu'il doit considérer dans l'accomplissement de sa fonction. Il est maintenant possible de comprendre l'importance de l'encapsulation pour la modularité fodorienne. En effet, l'encapsulation informationnelle des processus sous-jacents aux capacités psychologiques périphériques (i.e. les capacités perceptuelles, les capacités motrices et le langage) permettrait à ces derniers d'éviter les complications associées au problème du cadre. Autrement dit, si cela est vrai, l'encapsulation informationnelle d'un processus assurerait que les opérations de ce dernier soit computationnellement tractables.

Notons qu'en établissant la modularité d'un processus comme la condition de sa formalisation computationnelle, Fodor (1983, 2000) attribue à la modularité cognitive un rôle théorique similaire à celui que Marr (1982) accordait à sa propre forme de modularité. En effet, nous avons vu plus haut que, pour Marr, la nature modulaire du système visuel (i.e. l'autonomie fonctionnelle des différents sous-systèmes isolables responsables de la vision humaine) légitimait l'application de son analyse cognitive à trois niveaux aux différents processus du système visuel. De manière similaire, pour Fodor (1983, 2000), la nature modulaire de la périphérie de l'esprit (i.e. l'encapsulation informationnelle des structures sous-jacentes aux capacités cognitives périphériques) explique non seulement pourquoi les opérations des systèmes périphériques sont computationnellement tractables – ou, plus précisément, pourquoi elles ne font pas face au problème du cadre –, mais aussi, pourquoi il est légitime et fructueux de formaliser computationnellement les capacités psychologiques périphériques.

Dans la section suivante, nous nous limiterons à introduire la notion de modularité (i.e. la modularité darwinienne) impliquée la version adaptationniste de l'hypothèse de la modularité massive (aHMM). Pour ce faire, nous identifierons deux postulats modulaires primaires de la modularité darwinienne (i.e. les propriétés caractérisant ce type particulier de structure) et examinerons certaines inférences modulaires utilisées

par les partisans de l'aHMM pour identifier les postulats modulaires secondaires de la modularité darwinienne (i.e. les propriétés pouvant être exhibées par ces structures particulières).

1.2. La modularité darwinienne

Certains partisans de la psychologie évolutionniste[9] soutiennent que les composantes de l'architecture fonctionnelle de l'esprit humain sont principalement[10] les produits de la sélection naturelle. Selon cette thèse, les structures fonctionnelles de l'esprit humain seraient des adaptations psychologiques (i.e. des structures cognitives ou neurologiques ayant évoluées par sélection naturelle)[11]. Ceci n'implique toutefois pas que l'architecture de l'esprit humain soit uniquement constituée d'adaptations psychologiques. En effet, l'*adaptationnisme méthodologique* (voir Faucher et Poirier, 2009) adopté par les psychologues évolutionnistes est à distinguer du *pan-adaptationnisme*, selon lequel tous les traits des organismes seraient des adaptations (voir Gould et Lewontin, 1979). L'*adaptationnisme méthodologique* consiste plutôt en une méthode permettant de distinguer les traits ayant été sélectionnés de ceux ne l'ayant pas été (Thornhill, 2007, p. 33; voir également Andrews, Gangestad et Matthews, 2002). Ainsi, les partisans de la psychologie évolutionniste reconnaissent que l'architecture de l'esprit humain est constituée à la fois de structures fonctionnelles directement sélectionnées (i.e. les adaptations psychologiques), de structures non-fonctionnelles indirectement sélectionnées (i.e. les sous-produits (*by-*

[9] Nous reconnaissons que notre usage du terme « psychologie évolutionniste » réfère à un programme théorique particulier élaboré par un groupe restreint, mais proéminent, de psychologues évolutionnistes, souvent appelé l'École de Santa Barbara, et que ce programme n'est pas représentatif de la diversité des thèses et méthodes adoptées par l'ensemble des psychologues évolutionnistes (voir e.g. Dunbar et Barrett, 2007).

[10] « Outside of the operation of natural selection on our ancestors, there is no logical reason the brain should include any functionally organized elements beyond what random processes would produce » (Klein, Cosmides, Tooby et Chance, 2002, p. 307).

[11] Pour une comparaison de la notion d'adaptation psychologique avec la notion d'adaptation biologique, voir Hagen (2005, p. 156-157).

product) ou effets incidents (*side effects*) d'adaptations) et de structures (fonctionnelles ou non) préexistantes recrutées au service d'un nouvel effet psychologique bénéfique, et n'ayant pas subi l'action de la sélection naturelle pour l'accomplissement de ce nouvel effet (i.e. les exaptations[12]). Dans cette section, nous verrons que, malgré cette reconnaissance, au niveau descriptif, d'une pluralité de structures distinctes, certains psychologues évolutionnistes font le pari, qu'au niveau explicatif, ce sont les adaptations psychologiques qui doivent être les objets de la rétro-ingénierie des sciences cognitives.

Traditionnellement, les psychologues évolutionnistes se sont intéressés à un type particulier d'adaptations psychologiques humaines : les adaptations psychologiques propres aux *Homo sapiens* (i.e. en termes phylogénétiques, les *autapomorphies psychologiques humaines*). Toutefois, rien n'oblige qu'une adaptation psychologique humaine soit présente uniquement chez l'humain. En effet, certaines adaptations psychologiques humaines pourraient avoir été conservées par descendance à partir de traits ancestraux. Ainsi, en principe, un psychologue évolutionniste pourrait très bien s'intéresser aux adaptations communes aux primates, ou même aux mammifères, pour découvrir des propriétés nous informant directement sur les capacités psychologiques humaines. D'ailleurs, pour certains (e.g. Kaplan, 2002), les adaptations que nous partageons avec d'autres espèces sont la meilleure chance de trouver des données probantes sur les capacités psychologiques humaines. Néanmoins, vu la variabilité de l'organisation corticale entre les différents taxons (Preuss, 2001), il est plausible que certaines adaptations psychologiques humaines soient des *autapomorphies psychologiques humaines* (e.g. voir Barrett, 2012; Herrmann, Call, Hernández-Lloreda, Hare et Tomasello, 2007; Penn, Holyoak et Povinelli, 2008; Preuss, 2012). Cette précision nous sera utile, lorsque, dans la section 2.4, nous aurons à déterminer le statut phylogénétique de la structure psychologique spécialisée pour l'attribution d'états mentaux épistémiques.

[12] Pour une discussion de la notion d'exaptation de Gould et Vrba (1982) et une critique de la formulation de Buss, Haselton, Shackelford, Bleske et Wakefield (1998), voir Andrews *et al.* (2002).

La place centrale accordée aux adaptations par les partisans de la psychologie évolutionniste dans l'explication des capacités psychologiques s'explique par leur insistance sur les fonctions adaptatives. Selon Tooby et Cosmides (2005), une *fonction adaptative* représente les exigences fonctionnelles qu'une adaptation a dû satisfaire pour évoluer. Ces exigences fonctionnelles sont inférées grâce à l'analyse de tâche des hypothétiques pressions sélectives récurrentes (i.e. les problèmes adaptatifs) que l'adaptation en question a rencontrées dans son environnement adaptatif évolutionniste (EAE). De plus, l'identification d'une fonction adaptative permet d'informer la postulation d'un ensemble de structures sélectionnées qui interagissent de manière à résoudre des problèmes adaptatifs spécifiques (Tooby et Cosmides, 2005, p. 25). Cette formulation particulière de la notion de *fonction adaptative* intègre la notion de *fonction computationnelle*, conçue, selon Marr (1982), comme la spécification des conditions de réussite de la tâche de traitement de l'information qu'une structure doit effectuer, et la notion de *fonction d'effet sélectionné* (*selected effect function*), conçue, selon Williams (1966), comme la spécification des effets d'une adaptation pour lesquels cette adaptation évolua (voir Cosmides et Tooby, 1987, p. 287). Autrement dit, une fonction adaptative réfère aux effets d'une structure qui ont contribué à la propagation de cette dernière dans les populations ancestrales (Stotz et Griffiths, 2003). Si l'on accepte cette formulation, il s'ensuit que les effets psychologiques ne correspondant à aucune exigence fonctionnelle d'un problème adaptatif spécifique, ne sont, par définition, pas des fonctions et sont alors relégués au statut de simples effets incidents. En effet, les structures sélectionnées qui intéressent certains psychologues évolutionnistes sont non seulement des adaptations, mais sont aussi des structures spécialisées dans la résolution de problème adaptatif spécifique (nous revenons sur cette question particulière dans la sous-section 1.2.1).

Dans ce mémoire, nous appelons la thèse, selon laquelle l'architecture de l'esprit de l'*Homo sapiens* est composée principalement d'un grand nombre d'adaptations psychologiques dont la plupart sont spécialisées dans la résolution d'un problème

adaptatif spécifique, la version adaptationniste de l'hypothèse de la modularité massive de l'esprit (aHMM) (voir Barrett, 2006; Barrett et Kurzban, 2006; Pinker, 1997, 2005b; Sperber, 1994, 2002; Tooby et Cosmides, 1995a). Par exemple, selon Tooby et Cosmides (1995a) : « our cognitive architecture resembles a confederation of hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestors» (p. xiii-xiv).

Une des implications fondamentales et controversée de l'aHMM concerne l'existence d'un grand nombre d'adaptations psychologiques fonctionnellement spécialisées, non seulement à la périphérie de l'esprit, mais aussi au centre, où l'on retrouve les capacités cognitives de haut-niveaux responsables, par exemple, de la prise de décision, de la fixation des croyances et du raisonnement. Ainsi, contrairement à l'architecture modulaire de Fodor, l'aHMM soutient que des modules réalisent des capacités psychologiques évoluées de haut-niveaux. Cependant, l'aHMM n'est pas une application de la notion de module fodorien à l'entièreté du système cognitif (mais voir Bolhuis, Brown, Richardson et Laland, 2011). Au contraire, comme nous le verrons (sous-section 1.2.2), les modules darwiniens n'ont pas à posséder les différentes propriétés caractéristiques des modules fodoriens.

Comme nous le verrons dans cette section, l'aHMM implique l'existence d'une structure psychologique (i.e. le module darwinien) possédant des postulats modulaires primaires distincts des postulats modulaires primaires de la modularité fodorienne. En effet, le premier postulat modulaire primaire de la modularité darwinienne stipule que, pour être un module darwinien, une structure doit être spécialisée dans l'accomplissement d'un effet psychologique ayant contribué à la résolution d'un problème adaptatif (i.e être spécialisée pour une fonction adaptative). Le second postulat modulaire primaire stipule que cette structure doit avoir été directement sélectionnée précisément parce qu'elle accomplissait cet effet/fonction psychologique. Ainsi, selon Machery (à venir), un module darwinien est, à la fois, : 1)

une structure fonctionnellement spécialisée (i.e. un module fonctionnel ou anatomique) et 2) le produit direct de la sélection naturelle (i.e. une adaptation).

Dans le chapitre II, section 2.4, nous déterminerons si la région corticale sous-jacente à la capacité d'attribution d'état mentaux épistémiques (éTdE) exhibe les propriétés lui permettant d'obtenir le statut d'adaptation. À ce moment, nous reviendrons en détails sur les différentes conditions qu'un trait phénotypique doit satisfaire afin qu'on puisse lui attribuer avec assurance le statut d'adaptation. Pour l'instant, nous nous limiterons à préciser que les psychologues évolutionnistes sont tenus à une notion d'adaptation qui soutient que le statut d'adaptation est attribué à une structure si cette dernière satisfait un ensemble de critères indiquant que son parcours phylogénétique (historico-causal) fut directement influencé par la sélection naturelle (pour des discussions sur la valeur épistémologique de certains de ces critères, voir Andrews *et al.*, 2002; Brandon, 1990; Schmitt et Pilcher, 2004 ainsi que Simpson et Campbell, 2005). Cela n'implique toutefois pas qu'un module darwinien soit encore adaptatif dans l'environnement développemental, ni qu'il soit encore apte à effectuer la fonction adaptative pour laquelle il fut sélectionné.

Dans la sous-section 1.2.1, nous expliciterons le premier postulat modulaire primaire selon lequel les modules darwiniens doivent être fonctionnellement spécialisés au niveau cognitif (1.2.1.1) et au niveau neurologique (1.2.1.2). Ensuite, dans la section 1.2.2, nous expliquerons la raison pour laquelle, outre la spécialisation fonctionnelle évoluée, nous ne pouvons pas déterminer *a priori* les postulats modulaires des modules darwiniens. De plus, nous rejetterons l'idée que l'hétérogénéité des propriétés possibles des modules darwiniens entraîne la trivialisation de cette notion.

1.2.1. La spécialisation fonctionnelle évoluée

Certains psychologues évolutionnistes (e.g. Barrett, 2005, 2006; Barrett et Kurzban, 2006; Cosmides et Tooby, 1992, 1994, 1995; Pinker, 1997, 2005a, 2005b; Tooby et

Cosmides, 1995a, 1995b; Tooby, Cosmides et Barrett, 2005) et certains philosophes (e.g. Sperber, 1994, 2002, 2005), retiennent la notion de spécialisation fonctionnelle évoluée, plutôt que l'encapsulation, comme la propriété caractéristique des modules composant l'architecture de l'esprit humain. Le recours aux spécialisations fonctionnelles évoluées pour expliquer les capacités psychologiques humaines repose sur l'idée que les pressions sélectives de l'environnement seraient séparables en différents types particuliers de problèmes adaptatifs. De plus, ces types particuliers de problèmes adaptatifs seraient, en opposition à des problèmes généraux, des problèmes spécifiques. Par exemple, le problème des relations intra-espèces comprendrait différents problèmes adaptatifs spécifiques, notamment les relations parentales, l'accouplement, les négociations de statut hiérarchique, les relations de parentèle, les interactions inter-groupe et intra-groupe, la coopération et la compétition. Assumant la présence de plusieurs problèmes adaptatifs spécifiques, Cosmides et Tooby (1994; voir aussi Barrett et Kurzban, 2006; Carruthers, 2006; Gallistel, 2000) soutiennent que la sélection naturelle favorise l'évolution de solutions spécialisées propres à chaque problème adaptatif spécifique, plutôt que l'évolution de solutions génériques pouvant résoudre de manière satisfaisante plus d'un problème adaptatif (pour des critiques de cette position, voir Buller, 2005; Chiappe et MacDonald, 2005; Cowie et Woodward, 2004; Fodor, 2000, chap. 5; Lloyd, 1999; Samuels, 1998, 2006; Shapiro et Epstein, 1998; Stotz et Griffiths, 2003). Ainsi, selon Cosmides et Tooby (2005), si l'on découvre qu'une structure joue un rôle dans plusieurs capacités différentes, alors nous ne pouvons pas conclure que la structure en question est une adaptation ayant évolué pour l'accomplissement d'une capacité évoluée spécifique, mais devons considérer la possibilité qu'elle accomplisse une capacité plus générale.

La spécialisation fonctionnelle évoluée – ancestrale ou dérivée – est à distinguer de la spécialisation fonctionnelle proximale – cognitive ou neuronale. En effet, l'observation d'un module fonctionnel (i.e. spécialisation fonctionnelle proximale au niveau cognitif) ou d'un module anatomique (i.e. spécialisation fonctionnelle

proximale au niveau neurologique) n'est pas suffisante pour conclure de l'existence d'un module darwinien. En effet, une structure cognitive ou neurologique est un module darwinien si et seulement si sa spécialisation fonctionnelle est le produit direct de l'action de la sélection naturelle (Barrett, 2012). Le cas de l'apparente spécialisation fonctionnelle d'une région anatomique pour la lecture des mots (Cohen et Dehaene, 2004; Dehaene et Cohen, 2011) exemplifie clairement l'insuffisance de la spécialisation fonctionnelle proximale d'une structure comme critère de la modularité darwinienne. En effet, vu la récence phylogénétique de l'apparition de l'écriture, il est improbable que cette spécialisation soit le résultat de la sélection naturelle. Le module anatomique de la reconnaissance visuelle des mots ne serait donc pas un module darwinien, mais une exaptation recyclant/co-optant un système cortical phylogénétiquement antérieur spécialisé pour la recognition d'objet invariant (Dehaene et Cohen, 2011). En comparaison, Buss *et al.* (1998), considère que, sans l'influence de la sélection naturelle, une structure co-optant une adaptation antérieure n'est pas une exaptation, mais plutôt un nouvel usage non-fonctionnel (au sens biologique du terme) d'une adaptation préexistante (mais, pour une critique de cette formulation particulière, selon laquelle une exaptation doit avoir été sélectionnée, voir Andrews *et al.*, 2002). Notons que Brandon (1990, p. 172, note 14) soutient qu'une exaptation peut éventuellement être reconnue comme une adaptation si la persistance de cette exaptation dans une population est, en partie, causée par une sélection stabilisatrice (versus une sélection directionnelle ou une sélection perturbatrice). Selon cette interprétation, le module anatomique de la lecture des mots pourrait éventuellement devenir un module darwinien s'il subit au cours des générations futures l'influence d'une sélection stabilisatrice.

1.2.1.1. La spécialisation fonctionnelle évoluée au niveau cognitif

Au niveau computationnel (*sensu* Marr, 1982), un module darwinien effectue une

capacité psychologique évoluée spécifique. L'accomplissement de cette capacité psychologique évoluée doit, non seulement, avoir eu un impact positif sur la viabilité biologique (*fitness*) des variants ancestraux l'effectuant, mais aussi, avoir entraîné la sélection du trait phénotypique réalisant cette capacité spécifique[13]. De ce fait, comme nous l'avons vu plus haut, un module darwinien accomplit une fonction adaptative.

Au niveau algorithmique, si l'on choisit d'adopter une formalisation computationnelle, les procédures spécialisées des modules darwiniens accomplissent des transformations informationnelles dédiées à une capacité psychologique. Cependant, plusieurs raisons empêchent d'inférer avec certitude la nature exacte des algorithmes spécialisés à partir de la simple connaissance des pressions sélectives de l'EAE (pour une revue de ces difficultés, voir Machery, à venir). Bien qu'il soit possible d'utiliser les méthodes expérimentales de la psychologie cognitive et de la neuropsychologie pour dévoiler certains détails de ces procédures spécialisées, cette sous-détermination des opérations cognitives par les exigences fonctionnelles de l'EAE entraîne une ambivalence au sujet des implications algorithmiques de la spécialisation fonctionnelle d'une structure cognitive. D'un côté, Barrett (2012, p. 10735) reconnaît que les procédures génériques, telles que les règles d'inférences bayésiennes et les stratégies d'apprentissage statistique, peuvent décrire de manière satisfaisante la signature de certains modules darwinien. De l'autre, Pinker (1997, p. 31) affirme que les modules darwiniens « are defined by the special things they do with the information available to them, not necessarily by the kinds of information they have available ». En d'autres termes, Pinker (1997) considère que les modules darwiniens sont définis par leur spécialisation fonctionnelle, mais précise que cette spécialisation fonctionnelle doit être comprise en terme de la spécialisation de leurs procédures, et non comme une spécificité de leur domaine informationnel. En comparaison, Barrett (2005; voir aussi Barrett et Kurzban, 2006) propose de définir la spécialisation

[13] Un trait peut être adaptatif dans l'environnement ancestral et persister dans la population actuelle sans nécessairement avoir été sélectionné. Par exemple, suite à la modification de la fréquence d'un allèle ou d'un génotype indépendamment des mutations ou de la sélection naturelle (i.e. suite à une dérive génétique).

fonctionnelle des processus cognitifs comme la capacité de traiter un certain type d'information d'une certaine manière. En réponse à Fodor (2000), selon lequel une architecture computationnelle massivement modulaire ne peut expliquer les capacités psychologiques centrales humaines sans rencontrer divers problèmes (e.g. problème de l'input, problème de la globalité, problème de la sensitivité au contenu), Barrett (2005) développa une architecture modulaire « enzymatique » qui éviterait les écueils soulevés par Fodor (2000). S'inspirant du « global broadcasting » des informations (Baars, 1993; voir aussi Dehaene et Naccache, 2001) et du « constraint satisfying », cette architecture conçoit qu'un module puisse avoir accès, sans restriction, à un large éventail d'information (i.e. sans aucune spécificité d'accès (*access specificity*)) tout en demeurant spécialisé pour le traitement d'un domaine informationnel spécifique (i.e. avec une spécificité de traitement (*process specificity*)). Il suffit de savoir que dans ce modèle, si un processus est fonctionnellement spécialisé alors ce dernier possède nécessairement des restrictions formelles (i.e. des critères d'activation) spécifiant son domaine informationnel. Autrement dit, selon Barrett et Kurzban (2006, p. 630), la spécificité à un domaine serait une conséquence nécessaire de la spécialisation fonctionnelle évoluée de ce dernier. Cette formulation particulière de la spécificité à un domaine est conçue, par Carruthers (2006), comme une formulation unissant deux interprétations distinctes de la notion de domaine : l'interprétation en terme de format informationnel et l'interprétation en terme de problème adaptatif. En effet, la spécificité à un domaine informationnel et la spécificité à un problème adaptatif ne sont pas des propriétés équivalentes (voir, pour une position différente sur cette question, Barrett, 2009). La spécificité à un domaine, interprétée comme une spécialisation pour un type d'information, est un engagement ontologique à propos du format formel des entrées activant un module. En revanche, la spécificité à un domaine, interprétée comme une spécialisation pour une fonction adaptative, est un engagement ontologique à propos des effets qui ont provoqués la propagation phylogénétique de ce processus spécialisé.

Pour cerner la spécialisation fonctionnelle évoluée d'un module darwinien, il est important de distinguer le domaine propre d'un module de son domaine actuel (ou effectif) (Sperber, 1994). Le domaine propre représente les aspects stables de l'environnement pour lesquels la fonction cognitive fut *initialement* sélectionnée, alors que le domaine actuel représente le type d'input pour lequel le mécanisme cognitif spécialisé est *effectivement* utilisé. Cette distinction permet d'expliquer que les modules darwiniens n'ont pas à traiter exclusivement des informations relatives à leur domaine propre, car un module darwinien peut être co-opté par des informations évolutivement nouvelles (e.g. les voitures, les mots écrits) si ces dernières correspondent formellement au domaine propre du module. Ce phénomène de co-optation d'une structure spécialisée est d'ailleurs utilisé par les psychologues évolutionnistes pour expliquer les cas d'activation du gyrus fusiforme (le module spécialisé dans la détection et la reconnaissance des visages) chez des experts entraînés lors de tâches de reconnaissance d'oiseaux ou de voitures (Boyer et Barrett, 2005; voir également Duchaine, Yovel, Butterworth et Nakayama, 2006). De plus, on devrait s'attendre non seulement à ce que certains types ou formats d'information activent les modules darwiniens, mais aussi que cette activation soit associée à certains types d'activités et de circonstances écologiques ayant eu une importance substantielle et relativement uniforme pour la viabilité biologique des populations humaines ancestrales (Fessler et Machery, 2012). Par exemple, l'activation de certains schémas moteurs dépendrait non pas de la détection d'artefacts, mais de la détection d'artefacts manipulables (Boyer et Barrett, 2005). À ce facteur contextuel, s'ajoute le fait que la spécialisation fonctionnelle des modules darwiniens ne porte pas simplement sur les objets, mais sur certains aspects particuliers des objets. Par exemple, différents modules darwiniens seraient activés par différents aspects des visages : certains s'occupant d'identifier les individus, d'autres l'état émotionnel de ces derniers (Boyer et Barrett, 2005). De plus, certains modules darwiniens sont activés par plusieurs formats d'inputs. Par exemple, la physique intuitive applique des contraintes inférentielles à la fois sur les propriétés et les connections causales des

plantes, des animaux et des humains (Boyer, 2000). Alors que d'autres modules darwiniens possèdent des critères d'activation plus restreints. Par exemple, la présence d'états intentionnels serait inférée spontanément à condition qu'une autopropulsion et qu'une orientation vers un but soit simultanément détectées (Premack et Premack, 1995). De plus, le contexte écologique de la situation influence également quels modules darwiniens sont activés. Ainsi, la détection d'un animal peut, selon le contexte, activer soit les modules darwiniens responsables de la biologie intuitive (e.g. en contexte de prédation) ou bien ceux de la psychologie intuitive (e.g. avec un animal de compagnie) (Boyer, 2000). À ce stade, il est important de préciser qu'il est possible qu'un problème adaptatif soit composé de nombreuses sous-classes de problèmes, mais que ces dernières soient indistinguables du point de vue du domaine propre du module darwinien qui les résout (Barrett, 2009).

1.2.1.2. La spécialisation fonctionnelle évoluée au niveau neurologique

Malgré le fait que Barrett et Kurzban (2006, p. 640) soutiennent « that modularity in the sense of functionally specialized information processing can exist even in the absence of evidence of spatial localization from, for example, fMRI or lesion studies » et qu'ils « remain agnostic with regard to the way that functional specificity is implemented in the brain »[14], les modules darwiniens sont des entités neurologiques (e.g. systèmes, réseaux, aires, régions, circuits) (voir e.g. Barton et Harvey, 2000; Barrett, 2012; Machery, 2007; LeDoux, 2012). Une manière de représenter ce qu'implique la spécialisation fonctionnelle des modules darwiniens au niveau neurologique est de contextualiser cette spécialisation fonctionnelle dans le cadre de la différenciation cellulaire. Lors de l'ontogenèse du cerveau, comme pour tout organe du corps, les cellules embryonnaires totipotentes subissent un processus de

[14] Pour un argument selon lequel une description strictement fonctionnaliste des processus cognitifs est insatisfaisante, voir Buller (1993) ainsi que Buller et Hardcastle (2000, p. 319).

différenciation progressive menant à la formation de cellules, de tissus et d'organes spécialisés. Les neurones sont un type de cellules pluripotentes regroupant diverses classes et sous-classes de phénotypes neuronaux effectuant plusieurs fonctions (e.g. métaboliques, endocriniennes, immunitaires, sensorielles, motrices). Depuis Hartline (1938), il est reconnu que les fibres nerveuses accomplissent aussi des fonctions cognitives.

La spécialisation fonctionnelle est un principe important de la représentation et de l'encodage d'information sensorielle et motrice par le système nerveux périphérique (Olshausen et Field, 2004; Mahon et Cantlon, 2011). Bien que la spécialisation fonctionnelle des populations de neurones en périphérie du système nerveux soit plus évidentes, la question de l'extension, chez l'humain, de cette spécialisation fonctionnelle aux zones associatives et autres régions corticales associées aux capacités centrales (e.g. contrôle exécutif) est encore une question controversée (Mahon et Cantlon, 2011; voir, pour une revue de la spécificité fonctionnelle de la voie visuelle ventrale, Op de Beeck *et al.*, 2008; et, pour une revue des capacités de reconnaissance visuelle spécifique à un domaine d'information étant réalisée par des régions neurologiques spécialisées, Kanwisher, 2010).

Affirmer qu'une structure neurologique est spécialisée pour une fonction de niveau supérieur (e.g. une capacité psychologique) revient à établir « l'explication contextuelle » (*sensu* Craver, 2008) d'un rôle causal de cette composante de niveau inférieur (i.e. la structure neurologique). Autrement dit, une structure *P* pour être spécialisée pour fonction *F* doit être nécessaire à « l'explication constitutive » (*sensu* Craver, 2008) de la fonction *F*, sans toutefois avoir à être suffisante pour une « explication constitutive » complète. En effet, la spécification d'un rôle causal d'une structure sous-détermine la spécification des activités (*workings*) de cette structure (Bergeron, 2007, 2008). Idéalement, la spécification d'un rôle causal d'une structure neurologique doit tendre i) vers une description de la manière dont ce rôle causal particulier s'intègre (spatialement, temporellement et fonctionnellement) dans

« l'explication constitutive » de la capacité psychologique de niveau supérieur et ii) vers une élucidation empirique des activités réalisant ce rôle causal. De plus, spécifier le rôle causal d'une structure neurologique pour une fonction particulière n'implique pas que cette structure soit exclusivement dédiée à cette fonction. En effet, la spécialisation d'une structure neurologique pour une fonction implique minimalement qu'une structure participe à une fonction particulière plus fortement qu'à d'autres fonctions (voir Kanwisher, 2010). Dans le chapitre II (section 2.3), lorsque nous aurons à déterminer si la partie de la TdE traitant les états mentaux épistémiques (i.e. l'éTdE) est sous-tendue par une structure neurologique spécialisée, nous adopterons une formulation encore plus restrictive de la spécialisation fonctionnelle proximale qui satisfait, en plus du critère minimal de sélectivité neuronale, certains critères d'activation par différentes modalités à travers un ensemble de tâches différentes.

Pour illustrer brièvement ce qui est minimalement impliqué lorsque nous affirmons qu'une structure est fonctionnellement spécialisée, examinons le cas de la vision. Il est admis que le cortex visuel primaire (à distinguer d'une région du lobe occipital qui n'est pas une unité fonctionnelle, mais une unité structurelle) est spécialisé pour la fonction visuelle. En effet, bien que les activités du cortex visuel n'épuisent pas l'explication de la vision et ne sont pas exclusivement dédiées à la vision, il est admis que ces activités sont plus fortement recrutées pour la vision que pour d'autres fonctions et que cette participation est nécessaire à la fonction visuelle; ce qui permet d'affirmer que le cortex visuel est un module anatomique spécialisée pour la vision[15]. De même, selon ces critères, les rétines, les corps géniculés latéraux, les aires visuelles, les aires d'associations visuelle et certaines sous-composantes spécialisées pour la cognition spatiale des aires attentionnelles – en n'oubliant pas les circuits corticaux visuels et possiblement les colonnes corticales[16] qui composent ces aires –

[15] Cette inférence est effectuée sur la base de diverses critères anatomiques (e.g. architectonique, histologie et cytoarchitecture, connectivité neuronale) et fonctionnels (e.g. organisation des champs réceptifs, sélectivité neuronale) convergents.

[16] Le rôle causal des colonnes corticales est encore sujet de débat (Horton *et al.*, 2005; pour revue récente voir Kaas, 2012).

sont des structures anatomiques spécialisée pour la vision (i.e. chacune de ces structures, au niveau d'organisation qui lui correspond, contribue de manière sélective et nécessaire à la vision). Selon notre caractérisation de la modularité darwinienne (voir section 1.2), pour vérifier si ces différentes structures visuelles spécialisées sont des modules darwiniens, il suffirait en plus de démontrer que ces structures sont des adaptations ayant été sélectionnées pour leur spécialisation fonctionnelle respective.

Au niveau implémentationnel, les modules darwiniens ont peu de restriction concernant leur localisation anatomique. En effet, le fait d'être une structure neurologique à la fois spécialisée pour une capacité psychologique évolué et sélectionnée pour accomplir cette capacité, n'implique aucune hypothèse empirique particulière à propos de la localisation spatiale de cette structure (Cohen et Dehaene, 2004, Machery, 2007). Précisément, un module darwinien peut être distribué, multi-localisé ou même localisé discrètement (pour une analyse des différents types de localisation à différents types d'entités neurologiques, voir Mundale, 2003). Cependant, les opinions à ce sujet sont étrangement divergentes. D'un côté, certains auteurs (e.g. Anderson, 2010; Hardcastle et Stewart, 2002; Uttal, 2001) remarquent que les avancées neuroscientifiques, particulièrement celles recrutant la neuroimagerie, sont marquées par une accumulation de données expérimentales suggérant l'inexactitude des modèles modulaires du cerveau qui conçoivent que des régions anatomiques discrètes exhibent une spécificité fonctionnelle, notamment en ce qui concerne les fonctions cognitives centrales. De l'autre, certains auteurs (e.g., Barrett, 2012; Kanwisher, 2010) remarquent l'émergence d'un consensus à propos de la spécificité fonctionnelle avérée d'une multitude de régions corticales fonctionnellement spécialisées. Face à ce constat diamétralement opposé du statut théorique des disciplines neuroscientifiques utilisant la neuroimagerie, nous croyons pertinent de réaffirmer la conclusion de Mundale (2003), selon laquelle les principaux défis qu'auront à surmonter ces disciplines sont de nature taxonomique plutôt que technologique. En effet, à notre avis, la coexistence de deux interprétations

divergentes, chacune concluant à l'existence d'un consensus disciplinaire clair en faveur de deux interprétations en apparence diamétralement opposées, suggère, soit la présence d'un biais chez les deux camps dans la sélection des résultats représentatifs de l'état des neurosciences cognitives ou bien, plus probablement, la présence de certaines confusions conceptuelles empêchant ces deux camps de réaliser l'éventuelle compatibilité de leur interprétation des résultats de neuroimagerie. Notre opinion à ce sujet rejoint, au niveau méthodologique, celle d'Atkinson et Adolphs (2011), selon lesquels la combinaison de différentes méthodes neuroscientifiques (e.g. les enregistrements électrophysiologiques, les données de résonance magnétique fonctionnelles, les études de lésions, la stimulation magnétique transcranienne) permettra de dépasser les limites propres à chaque méthode et de produire des résultats plus fiables. De plus, nous croyons que la combinaison de différentes méthodes neuroscientifiques permettra aussi le développement d'inférences plus fines qui contribueront à l'atteinte de l'objectif de Bergeron (2007, 2008), selon lequel les attributions fonctionnelles se limitant à spécifier la participation (*cognitive roles/uses*) d'une structure à une activité cognitive doivent faire place aux attributions fonctionnelles spécifiant les activités (*workings*) des structures. Nous croyons que ces deux avancés méthodologiques permettront de mettre en évidence la plausibilité du phénomène de bifurcation (ou différenciation) ontogénétique des structures neurologiques spécialisées pour une capacité psychologique évoluée spécifique (voir Barrett et Kurzban, 2006, p. 439; Barrett, 2012, p. 10734). Ce scénario ontogénétique soutient que, plutôt que d'avoir des contributions causales recrutées pour plusieurs capacités (voir Anderson, 2010), une structure neurologique spécialisée pour une fonction adaptative subirait une différenciation développementale où des sous-ensembles de neurones distincts de cette structure spécialisée développeraient, chacun séparément, une spécialisation fonctionnelle proximale distincte en réaction à des demandes environnementales spécifiques, notamment suite à un entraînement par répétition (Jungé et Dennett, 2010). Notons que ce dernier modèle insiste sur la pluripotence initiale des populations neurones qui permettrait, par apprentissage, le

développement de modules anatomiques. De plus, il est important de spécifier que, bien que ce modèle de modularisation progressive en fonction de facteurs environnementaux spécifiques soit compatible avec l'aHMM, il ne l'est peut-être pas avec l'heuristique de décomposition mis de l'avant par les psychologues évolutionnistes qui insistent sur l'importance de l'analyse des fonctions adaptatives dans la détermination de la nature des structures fonctionnelles de l'esprit.

Au niveau ontogénétique, les structures neurologiques spécifiques à une capacité psychologique particulière (i.e. les modules anatomiques) sont les produits phénotypiques de procédures développementales. La modularité darwinienne implique que certaines procédures développementales furent sélectionnées en fonction de leur capacité à produire, de manière fiable, non pas une spécialisation fonctionnelle particulière, mais bien un *type* particulier de spécialisation fonctionnelle (Barrett et Kurzban, 2006; Barrett, 2006, 2007, 2012). En effet, puisque rien n'oblige qu'un module darwinien soit présent, dès la naissance, dans sa forme mature ou encore qu'il se développe indépendamment des influences environnementales, il est raisonnable de s'attendre à ce qu'un module darwinien, suivant sa norme de réaction développementale (*reaction norm*), puisse prendre plusieurs formes lors de son instanciation ontogénétique (Barrett, 2012). En effet, les procédures développementales des structures neurologiques spécialisées peuvent être sensibles à certaines influences contingentes, notamment par l'entremise de *patterns* locaux d'expression génétique (i.e. des *patterns* d'expression génétique propres à certaines structures anatomiques) (Barrett, 2012; voir également Geary et Huffman, 2002). Dans le chapitre II (section 2.4), nous reviendrons sur la question du développement ontogénétique des modules darwiniens lorsque nous aurons à déterminer si un module anatomique particulier fut effectivement sélectionné par la sélection naturelle pour développer, durant son ontogenèse, sa spécialisation fonctionnelle.

1.2.2. L'hétérogénéité de propriétés des modules darwiniens

Certaines critiques de la plausibilité de l'aHMM (e.g. Buller, 2005; Fodor, 2000, 2005; Lickliter et Honeycutt, 2003; Samuels, 2006) découlent d'un amalgame entre les propriétés cognitives et neurologiques impliquées par la modularité darwinienne et les propriétés typiquement associées à la modularité fodorienne. Ces dernières associent à la modularité darwinienne des propriétés qu'elle n'est pas requise de posséder (e.g. encapsulation, innéisme) (Barrett et Kurzban, 2006; Carruthers, 2006; Machery et Barrett, 2006). Face à cet abandon des propriétés associées à la modularité fodorienne, certains philosophes, tels que Fodor (2000, 2005), Prinz (2006) ainsi que Woodward et Cowie (2004), ont reproché à la notion plus libérale de module darwinien d'être une notion triviale, si cette dernière ne repose que sur la spécialisation fonctionnelle.

Selon Atkinson et Wheeler (2004), les sciences et les neurosciences cognitives évitent l'arbitrarité de la décomposition fonctionnelle, grâce à une rétroaction mutuelle entre les niveaux d'analyse et les niveaux d'organisation, permettant de contraindre empiriquement et théoriquement à la fois les décompositions fonctionnelles et les détermination de spécialisation fonctionnelle. Nous sommes d'accord que cette stratégie de « bootstraping » ordonné soit le seul moyen de se sortir de cet étang d'indétermination de principe de la décomposition fonctionnelle. Le point, souvent omis par certains critiques de l'aHMM (e.g. Prinz, 2006; Samuels, 2006), est que la modularité darwinienne ajoute, aux multiples contraintes proximales, des contraintes évolutives pour restreindre cette indétermination. En effet, comme nous l'avons spécifié plus haut (section 1.2), les modules darwiniens, en plus d'être des structures fonctionnellement spécialisées, doivent nécessairement être des adaptations. La modularité darwinienne implique ainsi, contrairement à la modularité fonctionnelle ou anatomique, les engagements empiriques découlant de son statut d'adaptation. En effet, selon Tooby et Cosmides (2005), la détermination de la fonction adaptative d'une structure, en plus d'informer si la structure que l'on examine est un module

darwinien ou une structure non-fonctionnelle, permet également d'inférer les propriétés respectives des modules darwiniens[17]. Cette inférence modulaire, permettant d'identifier les postulats modulaires secondaires des modules darwiniens (i.e. les propriétés que ces derniers peuvent exhiber), repose sur l'hypothèse de l'adéquation forme-fonction (*form-function fit*). Cette hypothèse affirme que la forme (ou la structure) d'une adaptation psychologique dépend de la fonction pour laquelle elle évolua. En raison de cette hypothèse d'adéquation forme-fonction des adaptations, les psychologues évolutionnistes considèrent qu'ils ont de bonnes raisons de s'attendre à ce que les modules darwiniens exhibent une hétérogénéité de propriétés neurologiques et cognitives diverses (Barrett et Kurzban, 2006; Barrett, 2012). Autrement dit, dans la lignée de Sperber (1994, p. 42), selon lequel les propriétés associées à la modularité darwinienne sont « a matter of discovery, not stipulation », si les modules darwiniens sont des structures neurologiques ou cognitives « façonnées » par la sélection naturelle pour résoudre un problème adaptatif spécifique, alors les propriétés attendues de ces modules dépendront, à la fois, de leur fonction adaptative particulière[18] (i.e les contraintes téléologiques) (Barrett et Kurzban, 2006; Machery et Barrett, 2006; Tooby et Cosmides, 2005) et de leur histoire évolutive particulière (i.e. les contraintes phylogénétiques) (Barrett, 2012). Nous réitérons donc qu'il est erroné d'inférer, à partir du constat de l'abandon des propriétés fodoriennes de la modularité cognitive (Fodor, 1983), que la modularité darwinienne équivaut alors à l'idée moins contentieuse d'une décomposition fonctionnelle en module fonctionnel ou anatomique. En effet, les modules darwiniens ne se réduisent pas à des modules fonctionnels ou anatomiques, car la modularité darwinienne réfère ultimement à des structures qui sont des adaptations, ce qui implique un ensemble complexe de conséquences empiriques

[17] Ce qui n'est pas sans poser problème (e.g. pour une discussion des difficultés associées à l'inférence des propriétés d'une structure à partir de son statut d'adaptation psychologique, voir Mameli, 2009).

[18] Par exemple, dans le modèle enzymatique de Barrett (2005), la spécialisation fonctionnelle d'une structure implique des hypothèses empiriques à propos du type d'entrée (input) qui active cette dernière (Barrett et Kurzban, 2006, p. 631).

fonctionnelles (i.e contraintes découlant de l'adéquation forme-fonction) et phylogénétiques (i.e contraintes provenant du parcours historique du trait).

En résumé, un module darwinien est une structure cognitive ou neurologique qui fut directement sélectionnée pour accomplir une capacité psychologique évoluée spécifique, c'est-à-dire, non seulement, une capacité ayant procuré un avantage au fitness des organismes ancestraux l'accomplissant, mais aussi ayant entraîné précisément la sélection de la procédure développementale permettant le développement ontogénétique de la structure spécialisée pour cette capacité psychologique évoluée. Cependant, cette entité neurologique n'a pas à effectuer exactement une seule capacité, ni à être dédiée exclusivement à une capacité, mais doit réaliser des procédures ayant été sélectionnées pour leur participation à une fonction adaptative « unitaire » selon le grain de description écologique qui importe au niveau d'explication (Brandon, 2005; Barrett, 2006). Ainsi, un module darwinien peut produire des effets exaptés pour lesquels il ne fut pas sélectionné (Sperber, 1994; Barrett et Kurzban, 2006). En effet, l'esprit/cerveau humain réalise plusieurs capacités psychologiques qui ne sont pas des effets fonctionnels de modules darwiniens. Par exemple, Kurzban *et al.* (2001) et Gil-White (2001) proposent tous deux que le racialisme serait un effet incident d'un module darwinien[19]. Ou encore Barrett (2012) et Dehaene et Cohen (2011) soutiennent que les mots écrits, malgré leur récence historique, sont traités par un module darwinien qui lui permet d'accomplir une fonction exaptée (i.e. la reconnaissance visuelle des mots) distincte de la fonction adaptative pour laquelle il fut hypothétiquement sélectionné (i.e. la recognition d'objet invariant). Cette fonction exaptée est alors un effet incident du module darwinien (Barrett et Kurzban 2006, p. 630)[20]. Par exemple, la main humaine, dont la

[19] Toutefois, ils ne s'accordent pas sur la nature du module darwinien entraînant cet effet incident. Kurzban *et al.* (2001) propose un module darwinien pour la détection d'alliances coalitionnelles, tandis que Gil-White (2001) propose un module darwinien pour un contexte de groupe différent, celui des ethnies.

[20] Si cette fonction s'avère être bénéfique pour l'organisme, ces effets incidents sont alors considérés comme exaptés (voir Andrews *et al.*, 2002).

fonction adaptative est probablement liée à l'utilisation d'outils (Marzke et Marzke, 2000) et au combat à mains nues (Morgan et Carrier, 2013), peut tout de même accomplir une panoplie de fonctions (e.g. applaudir) qui n'ont aucun lien avec ses hypothétiques fonctions adaptatives.

Plusieurs cas hypothétiques de modules darwiniens, possédant diverses propriétés, ont déjà été proposés comme explication de nombreuses capacités psychologiques centrales et périphériques (pour des revues, voir Duchaine, Cosmides et Tooby, 2001; Krill, Platek, Goetz et Shackelford, 2007). Par exemple, la mémoire déclarative et la mémoire procédurale (Sherry et Schacter, 1987), le module de choix des partenaires sexuels (Buss, 1989), le module du langage (Pinker et Bloom, 1990; Pinker 1994, 2003; Pinker et Jackendoff, 2005), les différentes heuristiques de jugement sous incertitude (Gigerenzer et Goldstein, 1996), le module de reconnaissance faciale (Kanwisher, McDermott et Chun, 1997; Kanwisher et Yovel, 2006), le module d'intuition des nombres (Spelke et Dehaene, 1999), le module de navigation spatiale et le module d'apprentissage de séries temporelles (Gallistel, 2000), le module de la peur (Öhman et Mineka, 2001), le module de détection de parentèle (Lieberman, Tooby et Cosmides, 2007) et la boucle phonologique (Aboitiz, Aboitiz et García, 2010). Le cas classique est celui du module darwinien de détection des tricheurs (Cosmides, 1989; Cosmides et Tooby, 1989, 1992; Cosmides, Barrett et Tooby, 2010; Gigerenzer et Hug, 1992)[21]. Précisément, selon Cosmides et Tooby (2005), le module darwinien de détection des tricheurs serait un sous-module d'une famille d'adaptations spécialisées pour les raisonnements en contexte d'échanges sociaux. Cette famille d'adaptations serait également composée de modules spécialisés dans la gestion de différents dangers intra-espèces spécifiques (voir Duntley, 2005).

La caractérisation conceptuelle de la modularité darwinienne que nous proposons est à distinguer des opérationnalisations particulières développées par les psychologues

[21] Malgré son statut paradigmatique, l'existence d'un module darwinien de détection des tricheurs est controversée (voir Buller, 2005, p. 171-172; Fodor, 2000; Gray, Heaney et Fairhall, 2003; Sperber, Cara et Girotto, 1995; Sperber et Girotto, 2003).

évolutionnistes. Elle tente de déterminer ce qu'un module darwinien devrait être en fonction de l'aHMM et d'établir des conventions et des repères utiles aux développement de notre compréhension scientifique de l'esprit/cerveau. Pour décomposer l'esprit/cerveau, on peut préférer, à la modularité darwinienne, des conventions conceptuelles inspirées d'autres disciplines. Par exemple, Bechtel (2003) favorise une notion au grain descriptif plus fin découlant des neurosciences cognitives : les opérations élémentaires. En comparaison, Boyer et Barrett (2005, p. 101-102[22]) considèrent que ce grain de description neurofonctionnel risque d'être trop fin pour identifier les adaptations psychologiques. La compatibilité de ces deux grains de décomposition fonctionnelle est une question ouverte (e.g. voir Anderson, 2010; Ritchie et Carruthers, 2010). Ultimement, la question du niveau de décomposition à adopter pour découper la nature à ses joints risque de dépendre des objectifs spécifiques des différentes disciplines. Par exemple, une structure neurologique fonctionnellement spécialisée qui n'est pas une adaptation (i.e. spécialisation fonctionnelle proximale) peut tout de même être un composant mécanistique d'un module darwinien et participer à la spécialisation fonctionnelle évoluée de ce dernier. Les difficultés potentielles d'une tentative de réconciliation des décompositions à un grain fonctionnel étiologique avec les décompositions à un grain fonctionnel proximal dépassent la portée de notre analyse (voir à ce sujet Bergeron, 2008, chap. 6). Nous remarquons tout de même qu'un champs de recherche en pleine expansion étudiant la connectivité entre les différentes aires cérébrales (i.e. leur architecture topologique) pourrait offrir un grain de description fertile pour observer les cas de spécialisation fonctionnelle évoluée du cerveau (e.g. voir Meunier *et al.*, 2010). Par exemple, certaines hypothèses de module fonctionnel évolutionnairement nouveau (*evolutionarily novel*) sont maintenant supportées par une approche comparative

[22] « in the current state of our knowledge of functional neuro-anatomy, it would seem that most functionally separable neural systems are more specific than the fitness-related domains, so that high-level domain specificity requires the joint or coordinated activation of different neural systems, and indeed in many cases consists largely of the specific coordination of distinct systems. » (Boyer et Barrett, 2005, p. 101-102).

évaluant les correspondances spatiales et temporelles de réseaux topologiques et fonctionnels entre les humains et les primates non-humains (e.g. Mantini *et al.*, 2013).

1.3. Contre une conception unifiée de la modularité en sciences cognitives

Dans ce chapitre I, nous avons examiné quatre usages distincts de la modularité développés respectivement par Marr (1976, 1982), Chomsky (1980, 1984), Fodor (1983) et certains psychologues évolutionnistes adoptant l'aHMM. Sans avoir présenté exhaustivement les justifications théoriques et empiriques offertes pour ces quatre usages de la modularité, nous espérons avoir suffisamment mis en évidence leurs postulats modulaires respectifs pour justifier la distinction de ces formes de modularité. Malgré une attention aux différents points communs entre ces quatre usages, ce premier chapitre est marqué par une insistance sur la disparité des usages de la modularité en sciences cognitives. En effet, nous ne proposons pas de caractérisation générale ou unifiée de la modularité en sciences cognitives. Nous nous sommes limités à identifier le rôle théorique joué par chacun de ces usages dans leur programme de recherche respectif et à souligner comment ces rôles théoriques influencent les postulats modulaires associés à chacun de ces usages. Par exemple, nous avons souligné comment, chez Marr et chez Fodor, le rôle théorique qu'ils accordent à la modularité (i.e. légitimer l'approche computationnelle) influence leurs différents postulats modulaires. De plus, nous avons vu comment la méthodologie adaptationniste de la psychologie évolutionniste contraint les postulats modulaires de la modularité darwinienne. En particulier, nous avons explicité le postulat primaire de la modularité darwinienne : la spécialisation fonctionnelle évoluée. L'analyse détaillée des implications de la spécialisation fonctionnelle évoluée au niveau cognitif et neurologique, nous a, entre autres, permit d'établir que la modularité darwinienne réfère ultimement à des structures fonctionnellement spécialisées qui sont aussi des adaptations. Plus précisément, nous avons vu que si une structure psychologique est

effectivement « façonnée » par la sélection naturelle pour résoudre un problème adaptatif spécifique, alors les propriétés attendues de cette structure particulière dépendront, à la fois, de sa fonction adaptative particulière (i.e les contraintes téléologiques) et de son histoire évolutive particulière (i.e. les contraintes phylogénétiques). Ainsi, la modularité darwinienne, malgré l'abandon des propriétés fodoriennes de la modularité cognitive, ne se réduit pas à la notion moins contentieuse de module fonctionnel ou anatomique, puisqu'elle implique, contrairement à la modularité fonctionnelle ou anatomique, des engagements empiriques découlant de son statut d'adaptation (dans le chapitre II, section 2.4, nous préciserons ces engagements empiriques en explicitant les critères qui permettent de déterminer le caractère adaptée ou évoluée d'une spécialisation fonctionnelle particulière).

CHAPITRE II

LA MÉTAREPRÉSENTATION D'INFORMATIONS MENTALES EST-ELLE SOUS-TENDUE PAR UN MODULE DARWINIEN?

> « Rather than asking whether the mind *must be* modular, they ask whether the mind *is* modular. And they start looking for accessible evidence that might help in deciding this issue. »
>
> (Mameli, M., 2001, p. 381)

2.1. Introduction

Dans le premier chapitre, nous avons établi qu'une structure cognitive ou neurologique est un module darwinien si et seulement si cette structure est, à la fois, 1) fonctionnellement spécialisée (i.e. un module fonctionnel ou anatomique) et 2) le produit direct de la sélection naturelle (i.e une adaptation). De plus, nous avons vu que ces deux postulats modulaires primaires de la modularité darwinienne déterminent, par l'entremise de l'hypothèse de l'adéquation forme-fonction, les autres propriétés possibles des modules darwiniens (i.e l'hétérogénéité des postulats modulaires secondaires). Dans ce second chapitre, afin de dépasser le débat portant sur la valeur des arguments de principe voulant établir la plausibilité ou l'implausibilité de la version adaptationniste de l'hypothèse de la modularité massive de l'esprit (aHMM), nous proposons de défendre l'aHMM en démontrant qu'une capacité psychologique de haut-niveau, la capacité d'attribution d'états mentaux épistémiques (éTdE), est sous-tendue par un module darwinien. Dans la section section 2.2, nous caractériserons l'éTdE comme une sous-capacité de la capacité de théorie de l'esprit (TdE). Dans la section 2.3, afin de satisfaire la première condition nécessaire de la modularité darwinienne, nous établirons l'existence d'une structure

neurologique spécialisée pour l'éTdE. Plus précisément, nous argumenterons que l'éTdE est la contribution causale d'une sous-région de la jonction temporo-pariétale droite (rTPJ) à la capacité plus générale de TdE (i.e. la capacité d'attribution, à soi-même et à autrui, de différents types d'états mentaux dans le but d'expliquer les comportements). Soyons clair, nous ne nous limitons pas à affirmer que l'éTdE nécessite la participation d'une sous-région de la rTPJ; nous soutenons, en nous basant sur un ensemble de données convergentes provenant de la psychologie cognitive, de la psychologie développementale, de la neuropsychologie et des neurosciences cognitives, qu'une sous-région de la rTPJ est non seulement plus fortement recrutée pour l'éTdE que pour d'autres capacités psychologiques (i.e. qu'elle est recrutée de manière sélective pour l'éTdE), mais aussi que la contribution de cette sous-région est unique à l'éTdE. Dans la section 2.4, afin de satisfaire la seconde condition nécessaire de la modularité darwinienne, nous examinerons s'il est plausible que la sous-région de la rTPJ spécialisée pour l'éTdE soit une adaptation. En résumé, nous défendrons, dans la section 2.3, que l'éTdE est sous-tendue par un module anatomique, pour ensuite, dans la section 2.4, évaluer s'il est plausible que ce module anatomique soit une adaptation ayant évoluée pour accomplir l'éTdE. Conformément à la caractérisation des modules darwiniens que nous avons proposée dans la section 1.2, nous considérons que si ces deux conditions nécessaires sont satisfaites, il est alors plausible que l'éTdE soit effectivement sous-tendue par un module darwinien.

Puisque, comme nous l'avons vu au chapitre I (sous-section 1.2.2), les modules darwiniens peuvent exhiber une hétérogénéité de propriétés (Barrett et Kurzban, 2006; Barrett, 2012), il ne nous sera pas nécessaire de résorber les tensions théoriques concernant la question computationnelle de la nature simulatoire ou théorétique de la TdE (voir sous-section 2.2.4) afin de défendre notre hypothèse selon laquelle une structure neurologique particulière est une adaptation spécialisée pour l'éTdE. En effet, ces deux conceptions de la TdE sont cohérentes avec l'idée qu'une sous-capacité

de la TdE (e.g. l'éTdE) est sous-tendue par un module darwinien. De plus, l'objectif ultime du chapitre II n'est pas d'approfondir spécifiquement notre connaissance de la TdE ou de ses sous-capacités, mais de démontrer, en support à l'aHMM, qu'il est plausible qu'une capacité psychologique centrale humaine (i.e. l'éTdE) soit sous-tendue par un module darwinien. Nous sommes conscients que même si cette dernière possibilité s'avérait la bonne, cela ne confirmerait pas la véracité de l'aHMM. Nous croyons toutefois que cette éventualité militerait fortement en faveur de la plausibilité de l'aHMM.

Cet argument repose sur la prémisse que l'attribution d'états mentaux épistémiques (l'éTdE) est une capacité psychologique centrale. Contre l'idée que l'éTdE serait une capacité psychologique périphérique (e.g. voir Samuels, 2006), ce qui viendrait diminuer considérablement le soutien de notre thèse en faveur de la plausibilité de l'aHMM[23], nous maintenons que l'éTdE est une capacité centrale. Toutefois, notre notion de « centralité » diffère de la notion de « centralité » mise de l'avant par Fodor (1983). Comme nous l'avons déjà mentionné (chapitre I, sous-section 1.1.5), selon Fodor, la fixation intelligente des croyances est la capacité qui caractérise le mieux les systèmes centraux. Selon lui, cette fixation intelligente des croyances doit être sensible à l'ensemble du système de croyances (i.e. aux propriétés globales du système de croyances) ce qui, par le fait même, la rendrait computationnellement intractable. Notre point est que l'éTdE, sans être sensible à l'ensemble du système de croyances, est tout de même une capacité psychologique centrale, car elle exhibe une sensibilité à différents contextes (*context-sensitivity*). De plus, cette sensibilité à différents contextes peut être computationnellement tractable si l'on reconnaît la possibilité que l'éTdE recrute des heuristiques frugaux (pour des précisions sur la frugalité computationnelle, voir Carruthers, 2006, p. 53-63).

Notre démonstration, dans la section 2.3, que l'éTdE est sous-tendue par une structure neurologique spécialisée n'implique pas que cette structure n'interagit pas de manière

[23] Sur ce point, nous tenons à remercier Pierre Poirier pour ses commentaires éclairants.

flexible avec d'autres structures neurologiques (voir sous-section 2.3.3, p. 70). Au contraire, il est possible de démontrer la sensibilité à différents contextes de l'éTdE en identifiant comment diverses sources d'informations peuvent influencer la prédiction des états mentaux épistémiques des individus. Par exemple, Koster-Hale et Saxe (2014) remarquent que, dans la majorité des expériences d'imagerie neurofonctionnelle supportant l'existence d'une structure neurologique spécialisée pour l'éTdE, la source informationnelle de prédiction « is not recent experimental history or trained associations, but rather a high level generative model of human thoughts and behaviors » (p. 9). De plus, dans ce même article, elles observent que les réponses de la région spécialisée pour l'éTdE sont non seulement modulées par les actions des individus, mais aussi par certains éléments du contexte social dans lequel sont effectuées ces actions (e.g. les normes sociales, l'origine sociale des individus).

2.2. Caractérisation de la théorie de l'esprit

Pour établir qu'il est plausible que l'éTdE soit la fonction évoluée d'un module darwinien, nous proposons d'examiner, dans la section 2.3, si les données scientifiques actuelles concernant cette capacité psychologique nous permettent de conclure qu'elle est effectivement sous-tendue par une structure fonctionnellement spécialisée. Pour ce faire, nous devons tout d'abord caractériser l'éTdE. Idéalement, pour bien cerner une capacité psychologique, nous devrions proposer une caractérisation rendant compte des quatre axes explicatifs proposés par Nikko Tinbergen (1963)[24] (i.e. l'axe mécanistique, l'axe ontogénétique, l'axe phylogénétique et l'axe de la fonction adaptative). Cependant, historiquement, la littérature scientifique concernant la TdE et ses sous-capacités fut principalement concentrée dans quatre disciplines, soit : la psychologie cognitive développementale, la psychologie cognitive sociale, la psychologie cognitive comparative (e.g.

[24] La classification de Mayr distinguant les causes proximales des causes ultimes possèderait certaines limitations (Ariew 2003; Downes 2005).

primatologie) et les neurosciences cognitives. Ainsi, dans la section 2.3, nous nous limiterons à une caractérisation de la TdE et de ses sous-capacités s'inspirant des explications mécanistiques et ontogénétiques[25].

Traditionnellement, le problème scientifique de la description et de l'explication de la capacité générale humaine à lire les pensées (*mindreading*) fut en grande partie lié au problème philosophique de la justification épistémique des connaissances acquises grâce à cette capacité mentale, particulièrement en ce qui concerne l'interprétation des pensées d'autrui (Nichols, à venir). La caractérisation de la lecture de la pensée que l'on adopte dans ce mémoire s'inscrit dans une tradition différente remontant à un article fondateur, rédigé par Premack et Woodruff (1978), qui posait alors la question : « Does the chimpanzee have a theory of mind ? ». Dans ce locus classicus, Premack et Woodruff définissent la TdE comme une capacité d'attribution d'états mentaux inobservables à soi-même et à autrui, mobilisée dans le but de prédire et expliquer les comportements de conspécifiques. Cette définition joue un rôle important dans notre caractérisation de l'éTdE, car elle spécifie que la TdE n'est pas qu'une simple anticipation du comportement à partir d'inductions empiriques ou d'associations. En effet, la TdE nécessiterait la postulation d'états mentaux comme causes sous-jacentes des comportements observés ou anticipés. De plus, cette caractérisation inférentielle de la TdE permet notamment de reconnaître que des organismes cognitifs non-linguistiques, tels que les jeunes enfants ou les animaux, peuvent aussi possiblement posséder, à des degrés divers, cette capacité d'attribuer des états mentaux. D'ailleurs, la question de la possession d'une TdE par des animaux non-humains fut récupérée par certains psychologues développementaux qui tentèrent de déterminer comment se développait cette capacité particulière chez les enfants normaux (e.g. Wimmer et Perner, 1983) et chez les enfants atteints d'autisme (e.g.

[25] La psychologie cognitive comparative se donne comme objectif de développer une caractérisation évolutionniste des capacités de cognition sociale que les différentes espèces de primates ont respectivement évoluées en insistant sur les différences et les similarités cognitives (e.g. Penn *et al.*, 2008; Vonk et Povinelli, 2006). Toutefois, certains primatologues considèrent que la caractérisation actuelle de la TdE demeure anthropomorphique (Barrett, Henzi et Rendall, 2007).

Baron-Cohen, Leslie et Frith 1985). Suite à l'approfondissement de la question du développement de la TdE par les psychologues développementaux, plusieurs raffinements furent apportés à la définition de Premack et Woodruff (1978).

Il est commun d'établir à la fois une distinction entre les différents types d'attribution, et entre les différents types d'états mentaux. Concernant les types d'attribution, nous distinguons l'attribution d'états mentaux à soi-même[26] de l'attribution d'états mentaux à autrui (e.g. voir Nichols et Stich, 2003). Pour ce qui est des types d'états mentaux, nous distinguons, suivant Baron-Cohen (1995), trois types principaux d'états mentaux : les états mentaux perceptuels (e.g. la direction du regard et les intentions gestuelles), les états mentaux volitionnels (e.g. les désirs, les préférences et les buts) et les états mentaux épistémiques (e.g. les croyances et les connaissances). De plus, nous distinguons les états mentaux des construits psychosociaux, comme les tendances dispositionnelles ou les traits de personnalité. Les attributions de construits psycho-sociaux sont regroupées sous le vocable « perception de la personne » (*person perception*) qui réfère aux attributions de traits mentaux plus stables et moins fugitifs que les causes mentales temporaires. Se basant sur les travaux de Leslie (1992, 1994), Baron-Cohen (1995) affirme que la TdE est réalisée par un module spécialisé pour inférer « the full range of mental states from behavior » (p. 51). Cette caractérisation de la TdE inclut les attributions d'émotions (Baron-Cohen, 2001) et les construits psycho-sociaux.

Comme nous l'avons mentionné plus haut, nous concentrons notre attention sur l'éTdE, que nous caractérisons comme la capacité d'attribution d'états mentaux épistémiques (i.e. la partie de la théorie de l'esprit (TdE) s'occupant de la représentation et de l'inférence des représentations épistémiques et de leur contenu informationnel) notamment parce que, comme nous le verrons dans la section 2.3, plusieurs données empiriques convergentes suggèrent que l'éTdE est une sous-capacité fonctionnellement dissociable de l'attribution des autres types d'états

[26] Stich et Nichols (2003) proposent que la capacité de s'attribuer des états mentaux dépend d'un module spécialisé pour le suivi de soi (*self-monitoring*).

mentaux. Notre hypothèse est que l'éTdE est une sous-capacité de la TdE, ce qui laisse la porte ouverte à la possibilité que d'autres composantes de la TdE s'occupent, spécifiquement ou en interaction avec d'autres composantes (voir sous-section 2.3.3), du traitement des autres types d'états mentaux, des émotions et des construits psycho-sociaux. Afin de caractériser l'éTdE comme une capacité de représentation du contenu des représentations épistémiques, aussi appelée la capacité de métareprésentation d'informations mentales, nous la distinguons de la capacité de représentation du contenu des représentations non-mentales, aussi appelée la capacité de métareprésentation d'informations non-mentales (e.g. la représentation de photographies ou de cartes), et de la capacité de représentation d'informations mentales non-épistémiques (e.g. la représentation d'émotions, d'état mentaux perceptuels ou volitionnels).

2.2.1. Le paradigme des tâches de fausses croyances

Jusqu'à récemment, les connaissances scientifiques sur les propriétés proximales de l'éTdE découlaient principalement d'un paradigme méthodologique développé en psychologie développementale. Ce paradigme s'inspire d'une idée de Dennett (1978), selon laquelle la meilleure manière d'étudier la compréhension d'un enfant à propos de la notion de croyance est d'examiner si ce dernier comprend qu'un individu entretient une fausse croyance. Selon cette idée, il serait possible de déterminer qu'un enfant comprend que les agents entretiennent des représentations du monde physique (i.e. qu'ils font preuve d'une intentionnalité de premier-ordre) en vérifiant si cet enfant attribue des états mentaux épistémiques à ces agents (i.e. qu'il fait preuve d'une intentionnalité de second-ordre). Pour vérifier si les enfants font preuve d'une capacité de métareprésentation d'informations mentales, Wimmer et Perner (1983) développèrent la célèbre tâche de Sally-Anne, qui avait pour objectif de vérifier si un participant (le sujet de la tâche) est capable d'anticiper correctement le comportement

hypothétique d'un agent fictif (une poupée nommée Sally) entretenant une fausse croyance. Les chercheurs présentaient aux participants deux poupées: Sally et Anne. Lors d'une courte mise en scène, on expliquait aux participants que Sally avait caché une bille dans un panier, mais que, par la suite, Anne avait, à l'insu de Sally, sortie la bille du papier en question pour la cacher dans son propre coffre. On demandait ensuite aux participants de répondre correctement à une question précise : où Sally cherchera-t-elle sa bille lorsqu'elle voudra la récupérer? Pour réussir cette tâche, un participant doit répondre que Sally cherchera sa bille dans le panier où elle l'avait initialement déposée et non dans le coffre de Anne où la bille se trouve finalement. Puisque Sally ignore la position réelle de la bille, on interprète la réussite de cette tâche comme suggérant non seulement que le participant a attribué à Sally une (fausse) croyance qui ne correspond pas à la (vraie) croyance que le participant a de la position de la bille, mais aussi qu'il a mobilisé cette attribution d'états mentaux épistémiques pour anticiper le comportement hypothétique de Sally. En d'autres termes, la réussite de cette tâche suggère que le participant possède une éTdE.

Afin de contrôler différentes variables (e.g. interaction avec le langage et/ou les fonctions exécutives; distinction entre les domaines du social, du mécanique et du biologique) et de vérifier la robustesse des résultats, la tâche de Sally-Anne fut par la suite développée en de multiples variations (e.g. test des Smarties, Perner *et al.*, 1989; vignettes non-verbales, Baron-Cohen, Leslie et Frith, 1986) constituant ce que nous appelons le paradigme méthodologique des tâches de fausses croyances. Après toutes ces variations expérimentales, il est possible d'identifier un résultat robuste découlant de ce paradigme de recherche : la majorité des enfants occidentaux de plus de 4 ans réussissent la tâche, alors que la majorité des enfants occidentaux de moins de 4 ans l'échouent. Ce résultat suggère que les *Homo sapiens* occidentaux développeraient, de manière relativement uniforme, une éTdE autour de l'âge de 4 ans. Notons qu'avec le récent développement de nouvelles tâches de fausses croyances non-verbales, il est maintenant clair que certains aspects de base (*core aspects*) de la TdE semblent se

développer beaucoup plus tôt que ne le laissait croire la piètre performance des enfants de moins de 4 ans aux tâches de fausses croyances verbales. Par exemple, Onishi et Baillargeon (2005) soutiennent, suite à l'étude des réactions visuelles de bébés lors d'une tâche non-verbale, que ces derniers démontrent, dès 15 mois, une compréhension implicite des états mentaux épistémiques d'autrui. Précisément, les bébés comprendraient qu'un adulte ignore la position d'un objet s'il est incapable de voir ce dernier. De plus, une étude récente de He, Bolz et Baillargeon (2011) indique que des enfants de 2 ans et demi peuvent réussir certaines variantes de la tâche des fausses croyances.

Bien qu'il soit admis que les enfant occidentaux font, de manière relativement uniforme, la démonstration de l'acquisition d'une éTdE autour de l'âge de 4 ans, il est maintenant reconnu que la distribution particulière des performance de ces enfants aux tâches de fausses croyances n'est pas uniquement attribuable au développement de l'éTdE. En effet, le pattern développemental robuste des performances aux tâches de fausses croyances découlerait d'un ensemble d'exigences attentionnelles et linguistiques supplémentaires associé à la fois aux demandes accessoires des tâches de fausses croyances et aux processus nécessaires au développement ontogénétique de l'éTdE (Bloom et German, 2000). En ce qui concerne les demandes accessoires des tâches des fausses croyances, Leslie, Friedman et German (2004) soutiennent que la réussite d'une tâche des fausses croyances autour de l'âge de 4 ans découlerait du développement d'une capacité de contrôle inhibitif, et non uniquement d'un changement conceptuel théorique. D'ailleurs, ce constat amène Leslie, German et Polizzi (2005) à décrire le mécanisme d'attribution des croyances, dans le contexte d'une tâche de fausses croyances[27], comme un mécanisme heuristique d'attention

[27] Leslie et Thaiss (1992) distinguent les attributions de représentations épistémiques en contexte des tâches de fausses croyances des attributions en contexte naturel (e.g. jeu symbolique), sur la base que les attributions de représentations épistémiques en contexte naturel découlent habituellement de l'observation directe du comportement de l'agent plutôt que d'une observation indirecte par l'entremise d'un récit verbal ou visuel. En effet, ces récits nécessiteraient, contrairement au contexte naturel, l'identification et la sélection par le participant des aspects pertinents de l'histoire correspondant à la perspective d'un agent fictif et non de la sienne.

sélective (*selection processor*), permettant de choisir entre différentes alternatives produites par un mécanisme spécialisé dans la manipulation d'états mentaux épistémiques innés (*theory of mind mechanism*). Sur ce sujet, le débat fait encore rage autour de la question de savoir si les enfants de moins de 4 ans échouent les tâches de fausses croyances parce qu'ils n'ont pas encore développé une éTdE, ou parce qu'ils n'ont pas encore développé les capacités inhibitrices permettant de répondre aux demandes accessoires des tâches de fausses croyances (nous reviendrons sur cette question dans les sous-sections 2.2.2, 2.3.2 et 2.3.3). Par exemple, Bloom et German (2000) remarquent que la difficulté de ces demandes accessoires est telle, que même une simple tâche d'inhibition non-représentationnelle est échouée par les enfants de 3 ans. Alors que Wellman, Cross et Watson (2001) soutiennent que malgré les différentes corrélations entre le contrôle inhibitif et les performances aux tâches de fausses croyances, on ne peut conclure que les enfants de moins de 4 ans qui échouent une tâche de fausses croyances possèderaient tout de même une éTdE.

Malgré que la question de savoir si ces capacités inhibitrices sont nécessaires au développement de l'éTdE ou si elles sont simplement nécessaire à la réussite d'une tâche de fausses croyances n'est toujours pas élucidée, pour Bloom et German (2000), le fait que la réussite d'une tâche de fausses croyances exige des ressources supplémentaires implique qu'un échec à une tâche de fausses croyances ne permet ni de distinguer entres les différentes causes spécifiques de cet échec, ni de conclure en l'absence d'une éTdE. Par exemple, il est probable que les autistes et les enfants de moins de 4 ans échouent pour des raisons différentes les tâches de fausses croyances (voir section 2.3.1, pour plus de détails à propos des performances des autistes aux tâches de fausses croyances). Pourtant, un échec à une tâche de fausses croyances ne permettent pas de distinguer ces deux phénotypes psychologiques pourtant très différents. De plus, Bloom et German (2000) remarquent que les comportements associés à l'éTdE ne sont pas épuisés par la capacité du sujet à représenter les fausses croyances d'un agent (voir aussi Samson et Apperly, 2010). Selon Bloom et German

(2000), ce dernier constat, ajouté au fait, mentionné plus haut, que la réussite d'une tâche de fausses croyances exige des ressources supplémentaires, justifierait l'abandon des tâches de fausses croyances comme critère décisif de l'acquisition ou non d'une éTdE. Nous verrons, dans la section suivante ainsi que dans la section 2.3.3, comment certains développements technologiques et méthodologiques permettent de contourner les limitations inférentielles du design expérimental des tâches de fausses croyances.

2.2.2. Nouvelles avenues

Récemment, nous assistons à plusieurs des transformations méthodologiques importantes permettant de combler certaines limites et faiblesses du paradigme des tâches de fausses croyances. Nous en soulignons trois qui contribuent particulièrement à l'argument de ce chapitre. Premièrement, au niveau de la psychologie développementale, on assiste, après un approfondissement quasi-exclusif du parcours ontogénétique de l'éTdE chez l'enfant, à une expansion des observations aux performances des adultes. Ainsi, Apperly, Samson et Humphreys (2009) avancent que les performances des adultes informent la psychologie développementale, puisqu'une description des processus matures de l'éTdE permet de déterminer à quel moment le développement de cette capacité est complété. De plus, ces derniers considèrent que les performances adultes informent différentes hypothèses à propos du rôle développemental respectif de capacités connexes, telles que les fonctions exécutives ou le langage. Par exemple, ils remarquent que les corrélations chez l'enfant, entre le développement de capacité inhibitrice et linguistique et les performances aux tâches de fausses croyances, indiquent probablement que les contributions de ces capacités connexes sont nécessaires au développement normal de l'éTdE, mais suggèrent qu'il est possible que ces contributions ne soient pas nécessaires aux performances de l'éTdE lorsque cette dernière est arrivée à maturité

(nous reviendrons plus en détails, dans la section 2.3.2 et 2.3.3, sur la question du rôle des fonctions exécutives dans le développement et les performances de l'éTdE). Pour l'instant, il suffit de remarquer que les coûts de traitement (*processing costs*) associés aux fonctions exécutives (e.g. la capacité de contrôle inhibitif), et découlant des demandes accessoires des tâches de fausses croyances, semblent être maintenus pour les performances adultes. Par exemple, Samson et Apperly (2010) considèrent que les demandes inhibitrices des tâches d'éTdE consistant à résister aux interférences de sa propre perspective ne sont pas des demandes accessoires des tâches, mais bien des demandes inhérentes. Contrairement au maintient de l'influence des capacités exécutives sur l'éTdE, les aptitudes grammaticales semblent ne plus influencer les performances adultes, alors qu'elles jouaient un rôle important dans le développement des performances des enfants (Apperly *et al.*, 2009). De manière générale, ces avancées méthodologiques permettent de déplacer l'attention aux conditions diachroniques d'acquisition développementale de l'éTdE vers les conditions synchroniques d'utilisation de l'éTdE lors des divers stades développementaux.

Deuxièmement, on assiste à un retour des mesures de l'exactitude (*accuracy*) de la compétence interpersonnelle d'attribution d'états mentaux, d'émotions, de trait stables, etc. En effet, traditionnellement en psychologie développementale, les performances ont servis d'indicateurs de l'atteinte ou non d'un stade développemental charnière. Une mesure de l'exactitude de la performance à une tâche implique le dépassement des limites des conditions expérimentales (e.g. les tâches de fausses croyances), permettant uniquement de déterminer la réussite ou l'échec d'une tâche. Par exemple, au niveau de la psychologie sociale, la prise en compte de l'exactitude des performances d'attribution ouvre un ensemble de possibilités inférentielles permettant, non seulement, de déterminer quels contextes et quels indices viennent influencer les activités de la TdE, mais aussi de quantifier l'ampleur de l'impact de ces influences internes et externes. Par exemple, Zaki et Oschner (2010), en plus d'insister sur l'importance d'une quantification de l'exactitude des attributions dans

certaines conditions, soulignent aussi l'importance de mesurer les variables altérant l'exactitude de ces attributions. À partir de ces variables, il serait possible d'inférer quels contextes (e.g. quels indices environnementaux) améliorent ou nuisent à l'exactitude de ces attributions d'états mentaux (e.g. voir Barrett, Todd, Miller et Blythe, 2005; Haselton et Funder, 2006). Cet aspect nous intéressera particulièrement, dans la section 2.4, lorsque nous évaluerons si la structure sous-tendant l'éTdE exhibe des indices de « design » adaptatif. De plus, les mesures d'exactitude fournissent des patterns d'erreurs pouvant nous informer directement ou indirectement sur les propriétés proximales des mécanismes d'inférences sociales, notamment en suggérant l'existence de biais. Par exemple, les mesures d'exactitude pourraient aider à trancher entre les différentes explications fonctionnelles concernant la nature simulatoire ou théorétique de la TdE (pour une discussion critique de la place de l'exactitude des attributions de la TdE dans ce débat, voir Zawidzki, 2008).

Troisièmement, dans certaines disciplines se concentrant traditionnellement sur les niveaux de description computationnels et algorithmiques (*sensu* Marr, 1982), on assiste à une intégration d'informations implémentationnelles, notamment celles découlant du développement des techniques d'imagerie neurofonctionnelle. Par exemple, Saxe, Carey et Kanwisher (2004) proposent une synthèse interdisciplinaire des observations de la psychologie développementale et des données d'imagerie par résonance magnétique fonctionnelle (iRMF) concernant la TdE. D'ailleurs, certains paradigmes expérimentaux recrutant des données d'iRMF permettent de contourner certaines limitations inférentielles du design expérimental des tâches de fausses croyances que nous avons soulignées dans la section précédente. Dans la section 2.3.3, ce type de données implémentationnelles nous permettra de préciser le rôle des capacités inhibitrices dans l'éTdE chez les sujets adultes neurotypiques. En ce qui concerne la TdE, cette tendance lourde[28] d'intégration des données comportementales et développementales avec les données neuroscientifiques supporte l'existence d'un

[28] Selon Koster-Hale et Saxe (2013), depuis le début du 21e siècle, plus de 400 études sur les régions corticales composant le système de mentalisation ont été effectuées.

réseau corticale spécifique à la TdE (e.g. Amodio et Frith, 2006; Gobbini, Koralek, Bryan, Montgomery et Haxby 2007; Mitchell, 2009; Saxe et Kanwisher, 2003; VanOverwalle, 2009; VanOverwalle, 2011, Zaki et Ochsner, 2013) se développant jusqu'à l'âge adulte (Saxe, 2010). Dans la section 2.3, un de nos objectifs sera de défendre l'hypothèse plus ambitieuse que ce réseau cortical spécifique à la TdE se dissocie anatomiquement et fonctionnellement en un ensemble de régions et sous-régions spécialisées dans le traitement de différents types d'états mentaux, notamment les états mentaux épistémiques.

2.2.3. Théorie-théorie et simulation

Deux approches dominantes proposent des types distincts de processus cognitifs pour expliquer la capacité générale de lecture de la pensée : la théorie-théorie (TT) et la théorie de la simulation (TS). De la manière la plus générale possible, la TT soutient, dans la lignée de Sellars (1956), qu'une structure interne « information-rich », implicite et/ou explicite, appelée par certains la « psychologie du sens commun » (*folk-psychology*), représente et spécifie comment les différents états mentaux interagissent ensemble pour informer et causer les actions. Ainsi caractérisée, la TT est une thèse qui spécifient exclusivement la nature des éléments et des interactions mécanistiques (*sensu* Tinbergen, 1963) expliquant la lecture de la pensée. Cette particularité explique que l'on retrouve actuellement, au sein de la TT, une cohabitation de diverses hypothèses développementales de la lecture de la pensée. Par exemple, Leslie *et al.* (2005) défendent une conception modulaire[29] de la maturation d'une théorie représentationnelle des états mentaux[30]. Gopnik et Schulz (2004), quant à eux, se basent sur des modèles formels d'inférences bayésiennes, pour expliquer une

[29] La notion nativiste de la modularité en psychologie développementale est à distinguer de la modularité darwinienne. Cette dernière n'étant commise qu'à un développement fiable (*reliable development*) (Barrett, 2006).

[30] Pour discussion des implications développementales de cette conception « modulaire », voir Scholl et Leslie (1999).

conception rationnelle, inductive et probabiliste de l'acquisition par apprentissage de théories intuitives causales. Il est admis que ces hypothèses distinctes s'inscrivent dans la TT puisqu'elles proposent toutes deux l'existence d'une structure interne « information-rich » pour expliquer, au niveau mécanistique, la capacité humaine à lire les pensée. Cependant, outre cette similarité mécanistique, il est important de remarquer que ces hypothèses proposent des processus ontogénétiques distincts pour expliquer l'acquisition d'une structure interne «information-rich ».

L'approche simulationniste soutient pour sa part que la lecture de la pensée est réalisée, au niveau mécanistique, par un processus ne nécessitant peu ou pas d'informations représentationnelles à propos du fonctionnement de ces états mentaux. Cette simulation mentale, nécessitant une similitude entre l'esprit interprétant et l'esprit à interpréter, découlerait d'une génération « off-line » des réactions potentielles de l'esprit interprétant dans la situation de l'esprit interprété. Au sein de l'approche simulationniste, on peut aussi répertorier plusieurs hypothèses distinctes pour expliquer la lecture de la pensée. Par exemple, Alvin Goldman (1989) défend une hypothèse selon laquelle l'introspection sert de base à un processus de projection personnelle délibérée permettant de comprendre la perspective d'autrui. Alors que Robert Gordon (1986, 1995, 2008) soutient qu'une simulation ne nécessite pas d'interprétation de soi préalablement à l'interprétation de l'autre, la simulation étant un processus de résonance (*mirroring*) informant directement nos interprétations et nos anticipations des comportements d'autrui. Précisément, Gordon (2008) avance que la simulation est un processus de compréhension directe ne nécessitant ni jugement à propos d'état mentaux, ni même la possession d'un répertoire classifié de ces derniers. Une autre hypothèse simulationniste de la lecture de la pensée découle des observations portant sur le système de neurones miroirs chez les primates (e.g. Gallese et Goldman, 1998; Gallese *et al.*, 2004) et insiste sur l'apport primordiale de la cognition motrice pour la cognition sociale (e.g. Gallese et Rochat, 2010).

Ces deux approches explicatives, autrefois considérées mutuellement exclusives

(Apperly 2008), sont intégrées par la majorité des récentes interventions au débat dans une approche hybride recrutant à la fois une théorisation et une simulation pour expliquer divers aspects de la lecture de la pensée (e.g. voir Carruthers, 2006, p. 176; Nichols et Stich, 2003; Goldman, 2006). Toutefois, une hypothèse récente, l'hypothèse de la pratique narrative (*Narrative Practice Hypothesis)*, proposée par Hutto (2007), remet en question à la fois les approches simulatoires et théorétiques. Cette hypothèse soutient que la lecture de la pensée est une pratique narrative nécessitant l'usage du langage et de scripts sociaux complexes. Elle propose que le but de la lecture de la pensée ne serait pas l'explication causale et la prédiction des comportements, mais plutôt l'interprétation des raisons et motifs d'une action. Cette interprétation reposerait sur la mobilisation de narratifs, littéralement d'histoires transmises culturellement, à propos des raisons et motifs d'agir spécifiques à certaines situations.

L'objectif de cette section n'est pas de trancher à propos de la nature simulatoire ou théorétique des structures spécifiques à la lecture de la pensée[31], mais plutôt d'établir le contexte théorique des distinctions nécessaires à notre caractérisation de l'éTdE. Nous nous limitons à établir que les aspects de la TdE qui nous concerne dans ce texte, sont ceux associés aux représentations d'état mentaux de « haut niveau ». Comme nous l'avons déjà mentionné, nous caractérisons l'éTdE comme la capacité humaine à représenter le contenu des état mentaux épistémiques (e.g. croyances, connaissances). L'éTdE est ainsi distinguée de la capacité de représentation des états mentaux de « bas niveau » (e.g. les émotions, les état mentaux perceptuels, les intentions gestuelles) et des représentations d'états mentaux volitionnels (e.g. les intentions, les préférences et les désirs). Notons que cette distinction bas niveau/haut niveau est orthogonale à la distinction simulation/théorie-théorie[32], car les attributions de haut niveau peuvent recruter de manière flexible à la fois des sous-processus

[31] Pour des discussions de cette question particulière, voir Apperly (2008), Gallese *et al.* (2004) ainsi que Saxe (2005, 2009).

[32] Pour une interprétation similaire de la distinction haut/bas niveau, voir Schultz (2011, p. 274).

effectuant des activités computationnelles catégorisées comme « simulatoire » et des sous-processus effectuant des activités computationnelles catégorisées comme « théorétique » (e.g. voir Zaki, Hennigan, Weber et Ochsner, 2010; Zaki, Weber, Bolger et Ochsner, 2009, concernant les attributions d'états affectifs). Dans la section suivante, notre objectif sera de déterminer si une structure particulière est spécialisée pour l'éTdE. Pour ce faire, nous examinerons des données comportementales, neuropsychologiques et d'imagerie fonctionnelle en faveur de l'existence d'une structure neurologique spécialisée pour la représentation du contenu des représentations épistémiques (i.e. la métareprésentation d'informations mentales).

2.3. Existe-t-il une structure spécialisée pour l'attribution d'états mentaux épistémiques ?

> « The discovery of the rTPJ, and the characterization of its functional specificity, serves as an existence proof that functionally specific cortical regions are not restricted to primary sensory and motor areas, or high-level perceptual regions, but can be found for at least one very abstract and high-level aspect of human cognition. »
>
> (Kanwisher, N., 2010, p. 4)

Tout d'abord, rappelons que, suivant notre caractérisation d'un module darwinien (chapitre I, section 1.2), l'hypothèse d'un module darwinien spécialisé pour l'éTdE ne serait pas affectée par les arguments infirmant la présence de certaines propriétés associées à la notion de modularité fodorienne. Par exemple, l'absence d'innéisme (e.g. Gerrans, 2003), d'automaticité (e.g. Apperly, 2006) ou d'encapsulation par rapport aux informations contextuelles (e.g. Currie et Sterelny, 2000) – telles que les normes sociales (e.g. Uttich et Lombrozo, 2010) ou les traits psychosociaux stables (e.g. Sripada, 2012) – ne s'opposent pas à la possibilité qu'une structure ait été « façonnée » par la sélection naturelle pour l'éTdE.

Dans cette section, nous adoptons une formulation de la spécialisation fonctionnelle

proximale inspirée par Kanwisher (2010). Nous soutenons que, pour pouvoir conclure à la spécialisation fonctionnelle proximale d'une structure neurologique pour l'éTdE, cette structure doit non seulement i) être plus fortement recrutée pour la représentation de représentations épistémiques que pour d'autres fonctions (e.g. la représentation de représentations non-mentales ou la représentation d'informations mentales non-épistémiques), mais aussi ii) s'activer pour tous les stimuli invitant l'attribution de croyances, vraies ou fausses (voir Saxe *et al.*, 2004).

La nature évasive de la confirmation d'une hypothèse scientifique ne nous permet que d'espérer amasser un nombre convaincant de données convergentes en faveur de l'hypothèse de la spécialisation fonctionnelle d'une sous-région de la rTPJ pour l'éTdE, et de comparer cette hypothèse avec les hypothèses alternatives expliquant ces mêmes données[33]. Ainsi, nous comparerons notre hypothèse d'une structure spécialisée pour l'éTdE (i.e. la métareprésentation d'informations mentales) avec le modèle alternatif le plus crédible qui propose une structure spécialisée pour la capacité générale de métareprésentation (i.e. une spécialisation pour toutes les formes de métareprésentation, incluant la métareprésentation d'informations non-mentales) (pour des arguments en faveur de cette dernière hypothèse, voir Egeth et Kurzban, 2009; Gerrans et Stone, 2008; Perner et Leekam, 2008; Stone et Gerrans, 2006). Pour ce faire, nous examinerons deux types d'évidences convergentes militant en faveur de la dissociation fonctionnelle de l'éTdE par rapport à la capacité générale de métareprésentation. Premièrement, dans la section 2.3.1, nous présenterons les déficits de performances aux tâches évaluant l'éTdE par les sujets atteints de certaines psychopathologies (e.g. autisme (Leslie et Thaiss, 1992), schizophrénie (Frith, 2004)) ou de certaines lésions corticales (e.g. Apperly, Samson, Chiavarino et Humphreys. 2004; Samson, Apperly *et al.*, 2004; Samson *et al.*, 2007, cependant voir, pour données divergentes Apperly, Samson, Chiaravino, Bickerton et Humphreys, 2007).

[33] « Given the observation, how credible is the theory? This depends on the plausibility of the theory, the plausibility of alternative modular theories, and the plausibility of single-process (i.e., non-modular) theories that are also consistent with the observation » (Sternberg, 2011, p. 183).

Deuxièmement, dans la section 2.3.2, nous présenterons une comparaison des temps de réaction à des tâches possédant des demandes computationnelles similaires, mais nécessitant le traitement de contenus informationnels distincts (e.g. German et Cohen 2010; Saxe et Kanwisher 2003). Après avoir présenté ces arguments en faveur de l'indépendance fonctionnelle de l'éTdE, nous défendrons l'existence d'un module anatomique spécialisé pour cette capacité. Pour ce faire, dans la section 2.3.3, nous démontrerons, en nous basant sur des données d'activation neurologique sélective (*neuronal selectivity*) robuste, l'existence d'un réseau corticale spécifique à la TdE (e.g. Amodio et Frith, 2006; Gobbini, Koralek, Bryan, Montgomery et Haxby, 2007; Mitchell, 2009; VanOverwalle, 2009; VanOverwalle, 2011; Zaki et Ochsner, 2013) se dissociant anatomiquement et fonctionnellement en un ensemble de régions et sous-régions spécialisées dans le traitement de différents types d'états mentaux (Carrington et Bailey, 2009; VanOverwalle et Baetans, 2009; Atique *et al.*, 2011). En particulier, nous nous attarderons aux études traitant de la participation de la région corticale appelée jonction temporo-pariétale droite (rTPJ) à l'éTdE (voir Atique *et al.*, 2011; Mitchell, 2008; Saxe et Kanwisher, 2003; Saxe et Wexler, 2005; Saxe et Powell, 2006; Scholz *et al.*, 2009). Dans la section 2.3.4, nous préciserons le rôle causal de la rTPJ à l'aide de données provenant d'études de lésions neurologiques.

2.3.1. Autisme : Déficit spécifique aux tâches de métareprésentation d'informations mentales ou aux tâches plus générales de métareprésentation?

Dans cette section, nous argumentons, contre Gerrans et Stone (2008; aussi Stone et Gerrans, 2006), en faveur d'une dissociation fonctionnelle entre la capacité plus générale de métareprésentation et la capacité plus spécifique de métareprésentation d'informations mentales. La première version de cet argument se retrouve dans la fameuse étude de Baron-Cohen, Leslie et Frith (1985) sur les autistes. Ces trois psychologues remarquèrent que les enfants autistes exhibent une dissociation de la

performance à une tâche non verbale de fausses croyances par rapport à des enfants d'âge mental équivalent atteints du syndrome de Down et neurotypiques. Précisément, Baron-Cohen, Leslie et Frith (1985) conclurent à un déficit spécifique de la capacité à représenter les états mentaux chez les enfants autistes. Dans le même ordre d'idée, les enfants autistes semblent performer normalement des inférences sociales lorsque ces dernières ne nécessitent pas d'attribution d'états mentaux, mais simplement une compréhension des dispositions comportementales (Baron-Cohen, Leslie et Frith, 2006). Cette particularité des performances métareprésentationnelles des autistes fut par la suite testée sur plusieurs types de représentations physiques et/ou publiques différentes (i.e. des métareprésentations d'informations non-mentales) à l'aide de tâches de fausses représentations physiques (e.g. photographies, cartes, signes). Ces tâches de fausses représentations physiques possèdent, en principe, une structure computationnelle similaire à la tâche des fausses croyances[34], à la différence qu'elles ne nécessitent pas la mobilisation de contenus informationnels mentaux. À partir d'une comparaison des performances à ces deux types de tâche, Zaitchik (1990) ainsi que Leslie et Thaiss (1992) observent une dissociation de la performance entre les enfants autistes et les enfants neurotypiques. Particulièrement, Leslie et Thaiss (1992) observent que les autistes réussissent mieux les tâches de métareprésentation d'informations non-mentales que les tâches de métareprésentation d'informations mentales, alors que les enfants neurotypiques exhibent un pattern inverse. Leslie et Thaiss (1992) infèrent à partir de cette divergence de la performance entre enfants autistes et enfants neurotypiques que ces deux types d'inférences sont réalisés par des processus cognitifs différents. Ils justifient cette hypothèse de la dissociation fonctionnelle, entre le traitement des métareprésentations et le traitement des métareprésentations d'informations mentales, sur la base de l'observation que la réussite, par les enfants neurotypiques, de la tâche de fausses croyances ne nécessite pas la réussite de la tâche de fausses photographies, et que la réussite, par les autistes,

[34] Pour des arguments contre l'équivalence des demandes computationnelles entre les tâches des fausses croyances et des fausses photographies, voir Perner et Leekham (2008) ainsi que Egeth et Kurzban (2009).

de la tâche des fausses photographies ne suffit pas à la réussite de la tâche de fausses croyances. En comparaison, Cohen et German (2010) concluent que les différentes expériences découlant du paradigme de recherche sur les habilités métareprésentationnelles des autistes, permettent d'affirmer que les autistes échouent généralement les tâches de fausses croyances et réussissent les tâches de fausses représentations physiques, mais que les enfants neurotypiques réussissent les deux types de tâches de manière relativement similaire.

Contre Leslie et Thaiss (1992), Egeth et Kurzban (2009) expliquent, à un niveau de grain moins fin, certains déficits de performance des autistes, par un déficit spécifique de la capacité générale de métareprésentation. Selon eux, la faiblesse de l'inférence de Leslie et Thaiss (1992), qui concluait à une dissociation fonctionnelle à partir d'une dissociation de la performance, découle de la possibilité que la divergence de la performance observée pourrait être causée par une asymétrie des demandes computationnelles entre les deux types de tâches (i.e. la dissociation entre la métareprésentation d'informations mentales et d'informations non-mentales serait un artéfact de performance causé par le fait que les tâches pour distinguer ces deux capacités ne seraient pas équivalentes). En conséquence, la dissociation de la performance pourrait indiquer, non pas un déficit lié aux contenus de la métareprésentation, mais un déficit lié aux difficultés représentationnelles plus élevées de la tâche des fausses croyances. Si, comme le soutiennent Egeth et Kurzban (2009), les tâches des fausses croyances s'avèrent plus difficiles que les tâches de fausses représentations physiques, alors l'échec, par les autistes, des tâches de fausses croyances, pourrait ne pas justifier l'inférence d'un déficit spécifique de l'éTdE. Notons que Saxe et Kanwisher (2003) observent que les réponses des sujets neurotypiques sont obtenues plus rapidement aux tâches de fausses croyances qu'aux tâches de fausses photographies, ce qui suggère, si l'on accepte l'hypothèse d'un module spécialisé pour le problème général de métareprésentation, que les tâches de fausses croyances ne seraient pas plus difficiles (dans la section 2.3.2, nous

présenterons d'autres données de temps de réaction concernant les sujets neurotypiques).

Gerrans et Stone (2008; aussi Stone et Gerrans, 2006), quant à eux, s'opposent à l'hypothèse d'une dissociation fonctionnelle entre la métareprésentation d'informations non-mentales et la métareprésentation d'information mentales, et misent sur un déficit spécifique de modules fonctionnels de bas-niveau pour expliquer les divergences de performances entre autistes et neurotypiques. Plus précisément, ils soutiennent que le déficit de performance des autistes aux tâches des fausses croyances s'expliquent par un déficit des capacités de bas-niveau alimentant la capacité générale de métareprésentation. Ainsi, les autistes auraient une capacité de métareprésentation intacte, ce qui expliquerait leurs réussites des tâches de fausses représentations physiques, mais ne pourraient obtenir les entrées (*inputs*) nécessaires aux inférences à propos des états mentaux épistémiques d'autrui. Toutefois, Adams (2011, 2012) démontre de manière convaincante que cette hypothèse particulière de Gerrans et Stone (2008) échoue à expliquer un cas particulier de déficit d'attribution d'états mentaux affectifs observé par Baron-Cohen *et al.* (1999). En effet, une mesure de contrôle effectuée par Baron-Cohen *et al.* (1999) s'assurait que les sujets autistes performaient de manière équivalente aux sujets neurotypiques lorsqu'on demandait d'effectuer une inférence sur les états mentaux affectifs des agents à partir d'un indice visuel (i.e. le regard)[35]. Cette tâche contrôle permet ainsi, selon Adams (2011, 2012), non pas d'évacuer complètement la possibilité d'un déficit de bas-niveau comme cause du déficit spécifique d'attribution d'états mentaux affectifs observé par Baron-Cohen *et al.* (1999), mais du moins, obligerait Gerrans et Stone à postuler l'existence de différents modules fonctionnels de bas-niveaux pour chaque modalité. Ce contre-argument potentiel irait, encore une fois selon Adams (2011, 2012), à l'encontre de

[35] Notons qu'une étude récente de Zürcher, Donnelly, Rogier, Russo, Hippolyte *et al.* (2013) observe que le traitement d'information mentales affectives et sociales à partir de visages peut être améliorée chez les autistes à l'aide d'une indication explicite de porter attention aux yeux. Cette observation pourrait expliquer pourquoi cette tâche contrôle particulière fut réussit par certains autistes considérant que la tâche contrôle de Baron-Cohen *et al.* (1999) insiste explicitement de s'attarder au regard.

l'objectif même de l'hypothèse défendue par Gerrans et Stone (2008), ces derniers voulant faire preuve de parcimonie en évitant la prolifération de modules fonctionnels. En effet, Gerrans et Stone (2008) proposent d'expliquer de manière plus parcimonieuse les patterns de performances comportementales et neurologiques associés à l'éTdE, par une interaction entre des processus de bas-niveau (e.g. reconnaissance des visages, détection et suivi du regard, détection d'agentivité) et plusieurs capacités de haut-niveau (e.g. fonctions exécutives, récursion, métareprésentation, mémoire de travail, etc.) (voir aussi Apperly et Butterfill, 2009). Cependant, Gerrans et Stone pourraient répondre à cette critique de Adams (2011, 2012) en indiquant que leur parcimonie se limitait à la postulation particulière d'un module fonctionnel spécifique à l'éTdE. Autrement dit, l'hypothèse défendue par Gerrans et Stone (2008) demeurerait valide et épuiserait les patterns de performance des autistes aux tâches de fausses croyances, si ces derniers acceptaient de multiplier les modules fonctionnels périphériques. Ce qui semble être une possibilité plus consensuelle.

Un autre problème est que Gerrans et Stone (2008) évacuent la possibilité que les performances des autistes découlent d'un déficit exclusif de la capacité plus générale de métareprésentation. Ils se basent pour soutenir ce point sur deux observations. Premièrement, ils remarquent que les symptômes des psychopathologies associées à un déficit des fonctions exécutives (e.g. syndrome de Down) ne s'apparentent pas aux symptômes présents chez les autistes. Deuxièmement, ils se basent sur les travaux de Griffith, Pennington, Wehner et Rogers (1999) pour conclure que les autistes ne semblent pas posséder de déficit des fonctions exécutives. Cette observation semble d'ailleurs supportée par le fait que les enfants autistes réussissent les tâches équivalentes de métareprésentation d'informations non-mentales (Leslie et Thaiss 1992). Cependant, Joseph et Tager-Flushberg (2004) suggèrent que les fonctions exécutives joueraient un rôle nécessaire - ils utilisent le terme « médiateur » - dans la réussite des tâches de TdE chez les autistes. Toutefois, ils remarquent : « it is not clear

from these data, nor from prior studies, whether executive functions are mainly important for performance on theory of mind tasks or whether they are more deeply involved in the conceptual developments that are necessary for a representational understanding of mind. » (Joseph et Tager-Flushberg, 2004, p. 11). Cette problématique, dépassant la portée de notre enquête, est directement attribuable aux limites inférentielles découlant du design expérimental des tâches de fausses croyances. Comme nous l'avons vu dans la section 2.2.1, les corrélations développementales et synchroniques, entre les performances aux différentes tâches de fonctions exécutives et aux différentes tâches de fausses croyances, ne peuvent révéler la nature des interactions entre ces deux habilités, car le design expérimental des tâches de fausses croyances ne permet pas de trancher entre un modèle où l'apport des fonctions exécutives est nécessaire, mais uniquement pour des demandes accessoires de la tâche (ces demandes influençant tout de même la performance des sujets), et un modèle où les fonctions exécutives jouent un rôle fondamental (synchronique et/ou développemental) dans l'interprétation des état mentaux. En somme, le rôle des fonctions exécutives pour la réussite ou l'échec, chez les autistes, des tâches des fausses croyances est encore une question ouverte. Nous verrons, dans la section 2.3.3, comment les données de neuroimagerie fonctionnelle permettent d'éclairer cette question particulière.

2.3.2. Temps de réaction des adultes aux tâches de métareprésentation mentale et physique

Face aux critiques de Egeth et Kurzban (2009) et de Gerrans et Stone (2008) concernant l'existence d'un processus de représentation spécifique aux représentations mentales sur la base d'une dissociation de performance chez les autistes (voir aussi Perner et Leekam 2008), on peut opposer des mesures de temps de réaction effectuées auprès d'adultes neurotypiques, lors de tâches recrutant des inférences

métareprésentationnelles comparant les performances en fonction du type de contenus représentationnels utilisé. Nous avons déjà mentionné dans la section précédente que Saxe et Kanwisher (2003) observent que les réponses des sujets neurotypiques sont obtenues plus rapidement aux tâches de fausses croyances qu'aux tâches de fausses photographies. Similairement, Cohen et German (2010) observent des temps de réaction plus faibles pour les jugements de croyances entretenues par un agent, à propos de la localisation d'un objet, à partir de questions formulées sous la forme de contenus mentaux (*belief probes*) que pour des questions formulées sous la forme de contenus représentationnels physiques *(map et arrow probes*). Contre les modèles architecturaux postulant l'existence d'un processus spécifique à la capacité générale de métareprésentation, Cohen et German (2010) considèrent que cette différence de temps de réaction supporte une interprétation architecturale postulant l'existence d'un processus spécifique à la capacité de métareprésentation des états mentaux. Notons que bien qu'il soit possible que cette plus grande rapidité découle effectivement de l'efficacité des inférences du processus spécifique à l'éTdE, il est aussi possible qu'elle découle d'une récupération plus rapide d'informations pertinentes activant le processus de l'éTdE, ou encore qu'elle découle d'un encodage d'informations sous la forme de contenus mentaux plus accessibles et saillants que l'encodage d'informations sous la forme de contenus représentationnels physiques (Apperly, 2010).

2.3.3. Activation neurologique sélective pour la représentation de différents types d'états mentaux

Il est important de répéter, avant d'approfondir la question de la sélectivité neurologique de certaines régions corticales à partir de données de neuroimagerie, qu'une hypothèse de spécialisation fonctionnelle proximale ne nécessite par la présence d'une hypothèse d'implémentation anatomique (Barrett et Kurzban, 2006; Bergeron, 2007). Cependant, nous concevons que les données de localisation

neurologique d'activités cognitives sont utiles, que ce soit par l'entremise de la « reverse inference » (Poldrack, 2006) ou de la « forward inference » (Henson, 2006), à la décomposition d'une capacité psychologique en opérations constituantes (Bechtel, 2002; Mundale, 2003; Friston et Price, 2011; Machery, 2014).

L'étude neuroscientifique de la capacité de la lecture de la pensée s'est développée autour de deux champs de recherches distincts observant des patterns d'activation neurologique robustes à deux types distincts de contenus mentaux. D'un côté, on observe des activations du cortex prémoteur, de la partie rostrale de la lobule pariétale et du sulcus temporal supérieur pour des tâches telles que la perception de la douleur, la perception du dégoût, la perception des émotions à partir d'expressions faciales (*social perception*) et la perceptions d'intentions gestuelles (*action understanding*) (e.g. Perlman, Vander Wyk et Pelphrey, 2010; VanOverwalle et Baetens, 2009; Zaki *et al.*, 2009). Ces activations robustes ont motivé l'hypothèse de l'existence d'un « système miroir » (Gallese, Keysers *et al.*, 2004) spécifique à la compréhension des actions et des comportements moteurs, activé par la présence de parties du corps humain (e.g. doigt, main, visage, membres) indépendamment de la modalité sensorielle, et à travers un ensemble de tâches distinctes (VanOverwalle et Baetens, 2009). D'un autre côté, on observe des activations du précuneus, des jonctions pariétales droite et gauche (rTPJ, lTPJ) et du cortex préfrontal médial (mPFC) pour les tâches de jugements d'intentions, d'émotions et de croyances (Amodio et Frith, 2006; Mitchell, 2009; Saxe et Kanwisher, 2003; VanOverwalle, 2009; VanOverwalle et Baetens, 2009). Ces activations robustes ont motivé l'hypothèse de l'existence d'un « système de mentalisation » (VanOverwalle, 2009; VanOverwalle et Baetens, 2009) ou « système d'attribution d'état mentaux » (Zaki *et al.*, 2010, Zaki et Ochsner, 2013) spécifique à l'interprétation inférentielle des intentions, des buts complexes, des croyances et même des traits stables, activés par différentes modalités (e.g. récits verbaux et non-verbaux, animations de formes géométriques) et à travers un ensemble de tâches sociales différentes, telles que l'interprétation des comportements,

des positions morales, des traits de personnalité et des traits de groupe (VanOverwalle et Baetens, 2009).

Un constat de VanOverwalle et Baetens (2009) est que ces deux systèmes sont dissociables fonctionnellement et n'interagissent pas pour accomplir les tâches auxquelles ils sont spécifiques. Toutefois, l'hypothèse de VanOverwalle et Baetens (2009), selon laquelle ces deux systèmes sont spécialisés dans l'interprétation de différents aspects du comportement (i.e. indices comportementaux corporels et indices comportementaux conceptuels), doit être nuancée par le fait que ces activations neurologiques sélectives à des types de contenus distincts, et indépendantes du contexte, découlent de situations expérimentales épurées représentant peu la complexité et la variété des stimuli traités dans les situations sociales naturelles (Zaki et Ochsner, 2009). Ainsi, une dissociation fonctionnelle de ces deux systèmes n'implique pas nécessairement que ces derniers n'interagissent pas. Par exemple, certaines tâches, telles que l'attribution d'émotions (Zaki *et al.*, 2009) ou l'interprétation d'indices sociaux ambigus (Zaki *et al.*, 2010) recrutent simultanément et de manière flexible ces deux systèmes. Ces dernières études suggèrent qu'il y aurait effectivement interaction entre ces deux systèmes dissociables fonctionnellement pour certaines fonctions plus générales pour lesquelles ils ne sont pas spécifiques.

En ce qui concerne l'hypothèse de la spécialisation fonctionnelle d'une structure corticale spécifique à l'éTdE, l'utilité de ces données d'iRMF découle du fait qu'elles permettent de dépasser certaines limites du design expérimental des tâches de fausses croyances en précisant le rôle des capacités inhibitrices dans l'éTdE. En effet, nous avions suggéré le fait que des données comportementales ne pouvaient démêler la corrélation entre les performances (réussite/échec) aux tâches de fausses croyances et les performances aux tâches de contrôle inhibitif à la fois chez les sujets neurotypiques (sous-section 2.2.1) et chez les sujets autistes (sous-section 2.3.1), L'identification neurologique du « système de mentalisation » permet de comparer les zones d'activation de ce réseau cortical aux zones activées lors de tâches de contrôle

inhibitif, et de vérifier si ces dernières concordent avec celles du « système de mentalisation ». Chez les adultes, les régions corticales respectives recrutées par chacune de ces deux fonctions ne correspondent pas, ce qui suggère que des structures distinctes réalisent l'éTdE et le contrôle inhibitif (Saxe *et al.*, 2004; voir, pour une méta-analyse de la question de la dissociation entre la capacité de métareprésentation mentale et les fonctions exécutives, VanOverwalle, 2011).

Nous avons établi, jusqu'à maintenant, qu'il existe un réseau neuronal, appelé « système de mentalisation », spécialisé dans l'attribution d'une variété d'informations mentales à partir de récits verbaux ou non-verbaux. Si l'on se rappelle, nous avions caractérisé l'éTdE comme une sous-capacité de la TdE : plus précisément, comme une capacité de représentation du contenu d'un type particulier d'états mentaux, soit les états mentaux épistémiques. Certaines données de neuroimagerie fonctionnelle et de neuropsychologie nous permettent d'affirmer qu'il existe, à un grain plus fin que celui dissociant le « système de mentalisation » et le « système miroir », une structure corticale spécifique à la représentation de ces états mentaux épistémiques (Carrington et Bailey, 2009, p. 2327). Cette hypothèse découle directement des travaux de Saxe et Kanwisher (2003) qui, suite à l'observation d'une plus forte activation des deux TPJ lors d'une tâche de fausses croyances que lors d'une tâche de fausses photographies, avancèrent initialement que les deux TPJ pouvaient être des régions spécialisées dans l'attribution de croyances à autrui. Cependant, cette hypothèse de spécialisation fonctionnelle pour l'éTdE fut ultérieurement restreinte à la rTPJ suite à une comparaison des activations avec la lTPJ lors de tâches de fausses croyances et de faux signes (Perner, Aichhorn, Kronbichler, Staffen et Ladurner, 2006). Cette hypothèse de la spécificité de la rTPJ à l'éTdE est également supportée par le fait que l'activation de la rTPJ est plus élevée lorsque des sujets lisent des récits incluant des états mentaux épistémiques, que lorsque ces récits incluent des informations sociales, telles que l'apparence d'un agent son appartenance culturelle, ou des informations mentales non-épistémiques, telles que la fatigue ou la faim (Saxe et Powell, 2006;

Saxe et Wexler, 2005). De plus, Saxe *et al.* (2004) considèrent que pour pouvoir conclure à la spécialisation fonctionnelle d'une région corticale pour l'éTdE, cette région doit non seulement être plus fortement activée pour le raisonnement à propos de la représentation du contenu d'états mentaux épistémiques que pour d'autres fonctions (e.g. la représentation d'informations représentationnelles non-mentales ou la représentation d'informations mentales non-épistémiques), mais aussi s'activer pour tous les stimuli invitant l'attribution de croyances, vraies ou fausses. À cet effet, Kobayashi, Glover et Temple (2007) ainsi que Carrington et Bailey (2009) maintiennent que l'activation des deux TPJ est indépendante de la nature verbale ou non-verbale des stimuli les recrutant.

Toutefois, une critique importante fut adressée à l'hypothèse de la spécialisation fonctionnelle proximale de la rTPJ pour l'éTdE suite aux observations d'une activation de la rTPJ lors de tâches de redirection de l'attention (Corbetta, Patel et Shulman, 2008; Decety et Lamm, 2007; Mitchell, 2008). Plus particulièrement, Mitchell (2008) avance à partir de ces résultats que la contribution causale de la rTPJ ne serait pas limitée à l'éTdE. À partir de ces nouvelles observations, trois interprétations possibles des données d'activation de la rTPJ peuvent cohabiter. Premièrement, il est possible que ces deux fonctions distinctes (i.e., métareprésentation épistémique et redirection de l'attention) recrutent deux sous-régions distinctes plus fines, mais indistinguables par la résolution spatiale de la méthode d'imagerie utilisée. Deuxièmement, il est possible que les deux tâches soit effectuées par la même région anatomique, cette région possédant tout simplement des rôles contextuels distincts en fonction de contextes d'activation différents. Ou encore, troisièmement, que les deux tâches soient en fait différents aspects d'une seule et même fonction commune plus générale et que les régions activées réalisent cette fonction plus générale. Conformément à la première possibilité, les résultats de Scholz, Triantafyllou, Whitfield-Gabrieli, Brown et Saxe (2009) démontrent que deux régions voisines, mais spatialement distinctes, de la rTPJ sont différentiellement

activées par les tâches de métareprésentation épistémique et les tâches de réorientation de l'attention. Ces résultats sont corroborés par Mars, Sallet, Schüffelgen, Jbabdi, Toni et Rushworth (2012) qui, à l'aide de mesures de la connectivité anatomique et fonctionnelle de la rTPJ chez des sujets au repos, concluent à l'existence d'au moins 2 sous-régions dans la rTPJ. Ces deux observations suggèrent qu'une sous-région de la rTPJ serait spécifique à l'éTdE (Saxe, 2010)[36].

Conformément à l'hypothèse de la spécialisation fonctionnelle proximale de la rTPJ à un type particulier d'états mentaux (i.e. les états mentaux épistémiques), Atique *et al.* (2011) observent que des sous-régions distinctes du rTPJ et du lTPJ seraient activées pour l'attribution d'intention et l'attribution d'émotions. Toutefois, Jenkins et Mitchell (2010) soulignent, en se basant sur leurs observations du mPFC, que cette stratégie de décomposition fonctionnelle, basée sur une différenciation des types d'attribution en fonction de différents types d'états mentaux, omet le fait que les patterns d'activation de certaines régions corticales du système de mentalisation pourraient représentés, non pas une spécificité à un type particulier d'états mentaux (e.g. les traits de personnalité stables pour le mPFC ou les croyances pour la rTPJ), mais plutôt une spécificité au degré d'incertitude ou d'ambiguïté du contexte d'attribution mentale (e.g. conspécifiques non familiers, situations nouvelles, comportements aux motivations ambiguës) et ce indépendamment du type d'états mentaux attribués. Dans une intervention récente, Koster-Hale et Saxe (2013) avancent l'hypothèse que le système neuronal de mentalisation serait constitué de plusieurs régions distinctes effectuant des prédictions spécifiques aux types d'états mentaux pour lesquels elles sont spécialisées. Cette hypothèse a le mérite de supporter la spécialisation fonctionnelle de différentes régions corticales du système de mentalisation à des types particuliers d'états mentaux, tout en restant cohérente avec les observations d'une

[36] Mais, voir Chater (2003, p. 168) : « Finding different 'hot spots' of neural activity for two different tasks cannot directly be taken as evidence for separate underlying cognitive machinery. It may be that the machinery is common for the two tasks, except for certain special components; or that all components are shared, but some are utilized more in one task, and some utilized more in the other task. ».

sensibilité de certaines de ces régions corticales au niveau d'ambiguïté des contextes d'attribution. Pour ce faire, cette hypothèse réinterprète l'activation sélective de certaines régions corticales lors de contextes ambigus d'attribution mentale en termes d'une sensibilité de ces régions aux aspects inattendus ou imprévisibles des stimuli. Ainsi, une région spécialisée dans la prédiction d'un type particulier d'état mental s'activerait plus fortement lorsqu'une situation ne correspond pas aux états mentaux anticipés que lorsque la situation observée correspond aux prédictions spécifiques de cette région.

En ce qui concerne les recherches de neuroimagerie liant autisme et TdE, nous nous limiterons à rapporter qu'aucun consensus actuel ne semble émerger en raison des résultats contradictoires observés (Saxe, 2010). Par exemple, la tâche classique de Heider et Simmerl (1944) fut administrée à des autistes dans plusieurs études différentes (e.g., Abell *et al.*, 1999; Bowler and Thommen, 2000; Klin, 2000). Au niveau comportemental, ces études démontrent que les sujets autistes utilisent des descriptions mentales explicites moins fréquemment et moins adéquatement que les groupes contrôles pour décrire les animations de formes géométriques (Castelli *et al.*, 2002). Ces observations comportementales ne sont toutefois pas entièrement corroborées au niveau neurofonctionnel, considérant que certaines études rapportent une hypoactivation de la rTPJ alors que d'autres études constatent une hyperactivation de la rTPJ (Saxe, 2010). Dans une étude récente, Lombardo, Chakrabarti, Bullmore et Baron-Cohen (2011) observent non seulement que la rTPJ des sujets autistes exhibe une réponse atypique comparativement aux sujets neurotypiques, spécifiquement lors de tâches nécessitant un traitement des états mentaux, mais aussi que cette réponse atypique est corrélé avec un déficit au niveau de la cognition sociale. Pour expliquer ce fonctionnement atypique de la rTPJ, Lombardo *et al.* (2011) proposent l'hypothèse que le développement normal d'une spécialisation fonctionnelle de la rTPJ serait perturbé chez les autistes. Notre intuition est que ces données apparemment contradictoires découlent de l'hétérogénéité des phénotypes neuronaux atypiques

satisfaisant les critères du diagnostic de trouble envahissant du développement. Ce diagnostic regroupe un ensemble varié de déficits sociaux (e.g. déficit de la TdE, déficit auto-référentiel, déficit du traitement des émotions, déficit du traitement des visages ou du regard, etc.) qui sont probablement explicables par une variété de phénotypes neurologiques distincts (Lombardo, Baron-Cohen, Belmonte et Chakrabarti, 2011; voir aussi Happé et Plomin, 2006). Par conséquent, bien que certains autistes exhibent un déficit spécifique de la TdE, ce déficit spécifique n'expliquerait pas tous les cas satisfaisant les critères diagnostiques de l'autisme (pour une discussion des insuffisances de l'interprétation de l'autisme comme une conséquence d'un déficit de la TdE, voir Baron-Cohen, 2010, p. 128; voir également Fisch, 2013).

2.3.4. Études de lésions

Parallèlement aux données de neuroimagerie, les études neuropsychologiques peuvent aussi informer la décomposition fonctionnelle de l'esprit/cerveau en examinant les déficits fonctionnels découlant d'une lésion neurologique[37]. Par exemple, dans le débat portant sur les décompositions fonctionnelles concurrentes à propos de la présence ou non de différents processus spécialisés sous-tendant les performances dans le raisonnement sur les contrats sociaux et les règles de précautions, plusieurs hypothèses maintiennent que ces deux tâches sont réalisées par un seul mécanisme cognitif plus général (i.e. spécifique aux raisonnements à propos des règles de permission, aux règles déontiques ou aux règles déontiques ayant des utilités subjectives). Contre ces trois différentes hypothèses d'une spécificité fonctionnelle plus générale, Cosmides et Tooby (2005) proposent que l'observation, par Stone, Cosmides, Tooby, Kroll et Knight (2002), d'une dissociation simple de la

[37] Contrairement aux inférences de modularité anatomique effectuées à partir de données de neuroimagerie, les inférences de modularité anatomique effectuées à partir de déficits fonctionnels découlant d'une lésion neurologique ne nécessitent pas d'engagement ontologique à propos de la localisation anatomique des capacités psychologiques.

performance entre des tâches de raisonnement sur les contrats sociaux et les règles de précautions, chez un patient lésé neurologiquement, milite en faveur de l'hypothèse d'un processus spécifique aux raisonnements sur les contrats sociaux. Ainsi, bien que des dissociations de la performance à ces deux tâches suggéraient déjà qu'elles soient réalisées par des processus distincts, il demeure qu'un déficit fonctionnel provenant d'une lésion neurologique permet de donner encore plus de poids à la décomposition fonctionnelle proposant un processus distinct spécifique à la détection des tricheurs.

Certaines données provenant d'études de lésions corroborent d'ailleurs notre hypothèse de l'existence d'une structure neurologique spécialisée pour la TdE se situant dans une sous-région ventrale de la rTPJ. Initialement, les déficits de la TdE étaient associés aux aires frontales. Par exemple, Stone, Baron-Cohen et Knight (1998) observent qu'une lésion du cortex orbitofrontal nuit aux performances de la TdE. Cependant, ces déficits de la TdE, suite à une lésion des aires frontales, sont toujours accompagnés de déficits comorbides à d'autres capacités de haut-niveau (Gerrans et Stone, 2008). De plus, on rapporte certains cas de dommage au mPFC associés à des performances intactes de la TdE (voir Cohen et German 2010). La nécessité de la contribution du lTPJ aux tâches des fausses croyances fut établie suite aux deux études de Apperly *et al.* (2004) et Samson *et al.* (2004) qui déterminèrent qu'un groupe de patients lésés à la lTPJ, réussissant toutes les tâches contrôles de fonctions exécutives et de mémorisation, non seulement échouaient une tâche de fausse croyance, mais ne démontraient même pas d'augmentation significative de leur taux de réussite à une tâche équivalente possédant des demandes inhibitrices et linguistiques plus faibles. La contribution de la lTPJ aux tâches des fausses croyances est cependant aussi nécessaire à la réussite des tâches de fausses photographies (Apperly *et al.* 2007), ce qui supporte l'hypothèse de Saxe (2009, 2010) que la lTPJ serait spécifique, non pas à l'éTdE, comme le serait la rTPJ, mais plutôt à la capacité générale de métareprésentation. À ce jour, aucun cas de lésion sélective à la rTPJ n'a été rapporté, mais cette hypothèse de Saxe (2009, 2010) implique qu'un patient ainsi

lésé présenterait exclusivement un déficit de l'éTdE, et non aux tâches de métareprésentation de stimuli physiques (e.g., photographie, cartes, signes).

En somme, après avoir défendu, contre Gerrans et Stone (2008) ainsi que Egeth et Kurzban (2009), l'indépendance fonctionnelle de la capacité spécifique de métareprésentation d'informations mentales par rapport à la capacité générale de métareprésentation, nous avons soutenu que cette capacité spécifique serait sous-tendue par une sous-région de la rTPJ. Plus précisément, grâce à un ensemble de données convergentes, nous avons établi que cette sous-région est spécialisée pour l'éTdE. En effet, cette sous-région i) est recrutée plus fortement pour l'éTdE que pour d'autres fonctions (e.g. redirection attentionnelle, représentation d'informations mentales non-épistémiques et métareprésentation d'informations non-mentales) ii) pour un large éventail de tâches et de types de stimuli (e.g. verbal et non-verbal). Bien qu'il soit clair que la contribution de la sous-région de la rTPJ n'épuise pas l'ensemble des observations concernant l'éTdE, les données recueillies jusqu'à ce jour démontrent qu'elle y joue un rôle causal spécifique. De ce fait, cette structure spécialisée pour l'éTdE satisfait le premier critère nécessaire à l'attribution du statut de module darwinien : la spécialisation fonctionnelle proximale d'une structure anatomique ou cognitive. Notons qu'il sera nécessaire de développer de nouveaux designs expérimentaux, si l'on veut étoffer notre spécification du rôle causal d'une sous-région la rTPJ dans l'éTdE et s'approcher d'une description exhaustive des activités constitutives de cette sous-région.

2.4. La région spécialisée pour l'attribution d'états mentaux est-elle une adaptation?

Dans la section 2.3, nous avons établi la spécialisation fonctionnelle proximale d'une sous-région de la rTPJ pour l'éTdE. Dans cette section nous examinerons si cette sous-région de la rTPJ satisfait le second critère nécessaire de la modularité

darwinienne : ce module anatomique doit être une adaptation (voir section 1.2). Autrement dit, nous examinerons la possibilité que la structure neurologique spécialisée pour l'éTdE soit effectivement le produit de la sélection naturelle.

Le statut d'adaptation d'un trait phénotypique est une hypothèse à propos du rôle de la sélection naturelle dans l'apparition phylogénétique d'une structure[38]. Cette hypothèse s'inscrit dans une tradition adaptationniste affirmant que la sélection naturelle peut modifier le comportement, les capacités psychologiques et les mécanismes neurologiques de manière à favoriser la « fitness » des organismes biologiques (e.g. voir Barkow, Cosmides et Tooby, 1992; Cosmides et Tooby, 2005; Gallistel, 2000; Krebs *et al.*, 1989; Shettleworth, 1998, 2000; Wilson et Daly, 1999). Cependant, cette tradition adaptationniste a fait face à plusieurs critiques de Gould et Lewontin (e.g. Gould et Lewontin 1979; Lewontin, 1979). De manière sommaire, Gould et Lewontin ont formulés deux importantes critiques épistémologiques de l'adaptationnisme. Premièrement, les adaptationnistes recruteraient des critères d'évidence inadéquats pour identifier la fonction et le statut phylogénétique d'un trait, notamment dans la détermination du statut d'adaptation (voir Andrews *et al.*, 2002; Schmitt et Pilcher, 2004 ainsi que Simpson et Campbell, 2005 pour des discussions sur la valeur épistémologique de certains de ces critères). Deuxièmement, les adaptationnistes ne prendraient pas suffisamment en compte la possibilité d'autres hypothèses que l'évolution par sélection naturelle pour expliquer les propriétés d'un trait phénotypique (pour une discussion sur l'importance de certaines contraintes évolutives pour le programme adaptationniste, voir Andrews *et al.*, 2002). À la lumière de ces deux critiques, nous examinerons, dans les sections 2.4.1 et 2.4.2, si nos critères d'évidence assurent l'attribution légitime du statut d'adaptation et, dans la section 2.4.3, si des scénarios évolutifs alternatifs peuvent expliquer l'existence d'une région corticale spécialisée pour l'éTdE.

[38] Notons que la sélection est alors dite *positive*, c'est-à-dire qu'elle participe au recrutement de nouvelles allèles (déjà présente dans une moindre mesure dans la population). En comparaison, la sélection *négative* assure le maintien d'un trait dans une population en éliminant certaines variations alléliques.

Avant d'examiner les critères d'évidence nous permettant de déterminer si la sous-section de la rTPJ spécialisée pour l'éTdE est une adaptation, il faut mentionner que deux conceptions de la notion d'adaptation cohabitent au sein de la tradition adaptationniste et que chacune de ces conceptions recrutent des critères d'évidence distincts. D'un côté, la conception historique soutient qu'une adaptation est un produit du processus d'évolution par sélection naturelle (Brandon, 1990). De l'autre, la conception « ingénierique » (*engineering conception*) avance plutôt qu'un trait qui exhibe un design fonctionnel complexe (*special design*) suggère la présence d'une adaptation (e.g. Cosmides et Tooby, 2005; Pinker et Bloom, 1990). Selon Lloyd (2012), ces deux conceptions distinctes, et possiblement incompatibles, de la notion d'adaptation, se retrouvent dans Williams (1966). Dans la section 2.4.1, nous évaluerons si l'on peut attribuer le statut historique d'adaptation à la sous-région de la rTPJ spécialisée pour l'éTdE. Dans la section 2.4.2, nous évaluerons si cette région répond aux critères d'évidence moins restrictifs de la conception « ingénierique » d'adaptation.

2.4.1. Conception historique

L'adaptationnisme et la phylogénie sont deux procédures complémentaires permettant de dévoiler le parcours phylogénétique (historico-causal) d'un trait phénotypique (Griffiths 1996; Thornhill 2007). Les analyses phylogénétiques s'intéressent à l'origine évolutive des traits et permettent d'établir les liens de parenté des organismes en fonction du degré de similarité de leurs caractères (phénétique) et de leur descendance à partir d'un ancêtre commun (cladistique). L'adaptationnisme s'attarde aux causes de la persistance phylogénétique des traits (e.g. types de sélection et valeur adaptative) et permet de distinguer les adaptations, des traits incidents (*by-products*) et des effets aléatoires (Buss *et al.*, 1998). Traditionnellement, les considérations phylogénétiques furent négligées par les psychologues évolutionnistes.

L'insistance de Tooby et Cosmides (1989, p. 182-186; 1992, p. 55) sur la primauté des considérations adaptationnistes à la fois comme source d'information sur les aspects comportementaux et cognitifs de l'esprit et comme cause unique des caractéristiques fonctionnelles des organismes fut probablement un facteur ayant contribué à cette négligence. Griffiths (1996) considère que cette limitation aux considérations adaptationnistes empêche la psychologie évolutionniste d'accéder à l'ensemble des méthodes empiriques qui permettraient de révéler le parcours phylogénétique d'un trait et de trancher à propos de son éventuel statut d'adaptation. Récemment, on assiste à un regain d'intérêt pour un type d'approche phylogénétique appelé la pensée homologique (e.g. Eastwick 2009; Ereshefsky 2007; Garcia 2010; Griffiths 2007; Love 2007; Matthen 2007). Cette dernière insiste sur une taxonomie des traits en fonction de leur ascendance commune et établit une relation d'identité en fonction de l'origine commune d'un trait[39]. Ainsi, certains auteurs (e.g. Matthen, 2007; Stotz et Griffiths, 2003) soutiennent que pour déterminer le problème adaptatif ayant permit le maintient d'une spécialisation fonctionnelle ou encore pour comprendre l'origine d'une innovation évolutive, il serait nécessaire d'examiner le contexte historique de son apparition (i.e. de connaître les variants ancestraux desquels l'innovation à permit une différenciation et ultimement un avantage adaptatif). Autrement dit, on ne pourrait étudier la nature d'une adaptation indépendamment de l'histoire des populations dans lesquels elle émergea et fut maintenue. D'autres (e.g. Barrett, 2012; Marcus, 2006; Thornhill, 2007) vont reconnaître l'importance de la pensée homologique en soutenant que les considérations adaptationnistes ne suffisent pas à dévoiler l'ensemble des propriétés des structures psychologiques ayant subies l'influence de la sélection naturelle puisque les structures sélectionnées d'un organisme héritent aussi de certaines propriétés par descendance à partir de traits ancestraux. En effet, l'inertie phylogénétique implique qu'une adaptation est toujours une modification d'une ancienne adaptation. Ainsi, selon Barrett (2012) : « This

[39] En comparaison, la pensée analogique catégorise les traits à partir de leur fonction adaptative (i.e. leur effet sélectionné) et la pensée phénétique à partir de leur similarité.

means that adaptations usually exhibit a mix of ancestral and derived features, which interact in their contribution to the adaptation's function » (p. 10735).

À ce sujet, certains auteurs (e.g. MacLean, Matthews, Hare, Anderson, Aureli *et al.*, 2012; Herrmann *et al.*, 2007) considèrent possible la reconstruction de l'état ancestral d'un trait à l'aide des outils de la psychologie comparative phylogénétique. D'ailleurs, selon MacLean *et al.* (2012), il serait possible d'identifier des changements évolutifs inter-espèces corrélés indiquant une sélection pour ce changement tout en contrôlant pour l'influence de l'inertie phylogénétique. Malheureusement, puisque ce type d'inférence statistique nécessite l'obtention d'un échantillon comprenant un grand nombre d'espèces exhibant le trait étudié (MacLean *et al.*, 2012; voir aussi Kaplan, 2002), il nous sera difficile d'obtenir des indications concernant le statut phylogénétique de la sous-région corticale spécialisée pour l'éTdE. En effet, il est plausible que l'éTdE (Call et Tomasello, 2008; Martin et Santos, 2014; Penn *et al.*, 2008) ainsi que la rTPJ (Saxe, 2006) soient des *autapomorphies psychologiques humaines* (voir section 1.2).

Devant l'éventualité que la région corticale spécialisée pour l'éTdE soit une adaptation propre à l'humain, nous tenterons de déterminer si effectivement la sélection positive contribua à l'apparition de cette structure corticale phylogénétiquement récente. Pour ce faire, nous nous inspirons des cinq critères, développés par Brandon (1990, p. 165-176), d'une explication idéale et complète d'une adaptation historique. Notons que Brandon (1990) prend l'exemple de la tolérance des plantes aux métaux lourds pour exposer cinq critères d'une explication adaptative complète. Nous considérons ces derniers suffisamment généraux pour qu'il soit valide de les reformuler et de les appliquer au cas de la rTPJ. Bien évidemment, l'attribution du statut d'adaptation à une structure psychologique n'a pas à satisfaire chacun de ces critères. Nous ne proposons donc pas de remplir chaque critère, mais plutôt d'examiner si des données préliminaires ou concluantes permettent d'indiquer qu'une sous-région de la rTPJ est plausiblement une adaptation pour l'éTdE plutôt

qu'une structure spécialisée ayant évoluée par chance ou le sous-produit (*by-product*) d'une autre adaptation.

1) Possédons-nous des données qui confirmeraient qu'une sélection des allèles responsables du développement ontogénétique de la région corticale spécialisée pour l'éTdE a historiquement eu lieu?
2) Possédons-nous une explication écologique du fait que les organismes possédant une région corticale spécialisée pour l'éTdE sont mieux adaptés que ceux ne possédant pas ce trait?
3) Possédons-nous des données confirmant l'héritabilité de la rTPJ?
4) Possédons-nous des informations à propos de l'évolution de la distribution génétique des *Homo sapiens* et des informations à propos de leurs différents environnements de sélection?
5) Possédons-nous des informations phylogénétiques concernant les traits ancestraux à partir desquels émergea la rTPJ?

Premièrement, possédons-nous des données qui confirmeraient qu'une sélection des allèles responsables du développement ontogénétique de la région corticale spécialisée pour l'éTdE a historiquement eu lieu? Ces données devraient à la fois nous indiquer que la structure psychologique spécialisée pour l'éTdE était mieux adaptée dans l'environnement de sélection que d'autres traits variants et nous confirmer que c'est cette augmentation de la « fitness » des organismes possédant cette structure spécialisée qui produisit la reproduction différentielle des variants ancestraux. Malheureusement, aucune donnée d'observation directe ne pourra être obtenue sur ce premier critère en raison de la nature non fossilisable du tissu cérébral. De plus, rien n'indique que nous serons en mesure d'obtenir des données historiques indirectes à ce sujet. En effet, il faudrait pouvoir établir le remplacement de variants ancestraux ne possédant pas de structure spécialisée pour l'éTdE par de nouveaux variants possédant cette structure spécialisée. Nous reviendrons sur cette question du

remplacement des variants ancestraux, lorsque nous évaluerons si nous sommes en mesure de satisfaire le cinquième critère.

Deuxièmement, possédons-nous une explication écologique du fait que les variants possédant une région corticale spécialisée pour l'éTdE sont mieux adaptés que ceux ne possédant pas ce trait? Au niveau écologique, différentes formes d'hypothèses tentent d'identifier les aspects environnementaux ayant « façonnés » l'évolution de la spécialisation corticale pour l'éTdE. Une première famille d'hypothèses, dite d'intelligence sociale, avance que la complexité des problèmes découlant des interactions sociales fut la principale pression sélective qui influença l'évolution de la cognition sociale humaine. Par exemple, les partisans de l'hypothèse de l'intelligence machiavélienne (une hypothèse spécifique de la famille des hypothèses d'intelligence sociale) soutiennent l'idée d'une course à l'armement de manipulation et de contre-mesure à ces tentatives de manipulation entraînant le développement de capacités de lecture de la pensée de plus en plus performante. Une seconde famille d'hypothèses, dite socio-écologiques, accorde une plus grande importance aux dynamiques sociales de coopération inter-individuelle pour expliquer l'apparition de la cognition sociale humaine. Pour les hypothèses socio-écologiques, la TdE est principalement un outil servant à la communication et à la transmission interpersonnelle d'information. Par exemple, la récente hypothèse socio-écologique de Tomasello, Melis, Tennie, Wyman et Herrmann (2012) inclut à la fois des contextes d'interaction coopérative intra-groupe, telles que les situations de chasse collective, et des contextes d'interaction compétitive inter-groupe, tels que les conflits tribaux, comme pressions sélectives ayant influencées la formation de la cognition sociale humaine.

Au niveau neurologique, on ne peut ignorer la corrélation de l'encéphalisation des primates avec la complexité des relation sociales. En effet, les cerveaux humains sont environ trois fois plus volumineux que ceux des chimpanzés, et ce malgré le fait que la taille absolue des régions sensorielles et motrices soit comparable à celle des grands singes (Preuss, 2011). Des données récentes, s'inscrivant dans l'hypothèse du

cerveau social (*social brain hypothesis*), semblent indiquer que l'augmentation du volume du néocortex des primates est liée au degré de complexité des relations sociales et non simplement à la taille des groupes comme on le croyait initialement (Dunbar et Shultz, 2007). Cependant, plusieurs questions persistent au sujet des demandes cognitives associées à ces relations sociales et des aspects écologiques qui rendirent ces relations sociales si avantageuses. Récemment, Dobson et Sherwood (2011) ont proposé l'hypothèse que la perception des émotions exprimées par les visages jouerait un rôle fondamental dans l'encéphalisation du néocortex des singes de l'Ancien Monde (Catarhiniens). Malgré cette plausible influence de la sélection naturelle sur l'évolution des structures corticales de certains primates, il reste à savoir si les nouvelles capacités psychologiques humaines (i.e. les *autapomorphies psychologiques humaines*) découlant de l'encéphalisation du néocortex humain sont rendues possible grâce à l'augmentation du nombre de connections neuronales ou grâce à l'apparition de propriétés particulières aux circuits neurologiques humains (pour une revue de la question voir Buckner et Krienen, 2013). En ce qui concerne la rTPJ, l'augmentation disproportionnelle du volume des lobes temporaux serait principalement associée à une augmentation de la matière blanche, ce qui indique une augmentation à la fois de la connectivité interne des lobes temporaux et de la connectivité de ces derniers avec les lobes frontaux et pariétaux (Schenker, Desgouttes et Semendeferi, 2005).

Troisièmement, possédons-nous des données confirmant l'héritabilité de la rTPJ? Hughes et Cutting (1999) effectuèrent la première étude de jumeau tentant de déterminer l'héritabilité des performances aux tâches de TdE. Ces derniers avancent que les performances de la TdE sont fortement héritables. Plus récemment, Xia, Wu et Su (2012) observèrent une association entre certains polymorphismes d'un seul nucléotide (*single nucleotide polymorphism*) et les performances à des tâches de TdE. Pour qu'il y ait sélection d'un variant phénotypique, ce variant doit être héritable. Ces observations suggèrent qu'il est possible que la sélection naturelle ait agi sur les

différentes structures psychologiques responsables de la TdE puisque les variations de performance de cette capacité sont fortement héritables. En ce qui concerne précisément, la sous-structure de la rTPJ spécialisée pour l'éTdE d'autres études plus précises seront nécessaires avant de pouvoir se prononcer.

Quatrièmement, possédons-nous des informations à propos de l'évolution de la distribution génétique des *Homo sapiens* et des informations à propos de leurs différents environnements de sélection? Il est possible de déterminer, à l'aide de différentes méthodes en génétique évolutionniste, si des régions du code génétique humain furent récemment ou sont encore sous l'influence d'une sélection directionnelle. Par exemple, les *loci* génétiques associés à la résistance à la malaria, la pigmentation de l'épiderme ou la tolérance au lactose sont des candidats au statut d'adaptation spécifique de l'*Homo sapiens* (Nielsen, Hellmann, Hubisz, Bustamante et Clark, 2007). Bien qu'il y ait très peu de chance qu'une fonction cognitive découle d'un locus génétique unique (voir Ramus, 2006), la détection d'une sélection positive pour le gène *FOXP2* démontre qu'un gène associé à des fonctions cognitives peut avoir été sélectionné. En ce qui concerne la rTPJ, à notre connaissance, aucune étude génétique associé à cette région corticale ne permet de conclure qu'il s'agirait du produit d'une sélection directionnelle.

Cinquièmement, possédons-nous des informations phylogénétiques concernant les traits ancestraux à partir desquels émergea la rTPJ? Sur ce point, il est difficile d'établir l'histoire évolutive de la structure neurologique particulière qui nous intéresse, entre autres, parce qu'il est difficile d'en établir une caractérisation moléculaire et neuro-cellulaire. De manière peu encourageante, même si nous arrivions à développer une caractérisation moléculaire des opérations de la rTPJ, selon Bickle (2008), les leçons évolutionnistes des travaux d'Éric Kandel concernant la caractérisation moléculaire de la capacité de consolidation mnésique suggère que les neurosciences moléculaires et cellulaires risquent, pour l'instant, de ne pas pouvoir nous informer sur les *autapomorphies psychologiques humaines*. En effet, Bickle

(2008) soutient que les traits psychologiques qui ont, à la fois, un impact causal sur la valeur d'adaptation de leur porteur et la possibilité d'être étudiés par les neurosciences moléculaire sont, pour l'instant, uniquement des traits relativement stables phylogénétiquement. Bickle (2008) explique ce constat par le fait que les transformations informationnelles du système nerveux sont sous-tendues par les taux de modification des potentiels d'action des circuits neuronaux, que ces taux de modification découleraient de processus relevant du métabolisme de base (e.g. les processus régulant les échanges d'énergie et les modifications intracellulaires) qui, à leur tour, dépendraient directement de protéines et d'enzymes dont le rythme de modification phylogénétique est lent. Le projet de Griffiths (2006; 2007; voir aussi Brandon, 2005) d'articuler les décompositions fonctionnelles des traits psychologiques à partir de la notion d'homologie est, à notre avis, une réponse compatible avec cette limitation scientifique des neurosciences moléculaires. En effet, en identifiant la spécificité fonctionnelle d'une structure homologue, un homologue fonctionnel, on peut alors étudier un type naturel plus stable phylogénétiquement. À notre avis, les homologues fonctionnels sont des traits psychologiques qui pourraient offrir une meilleure « commensurabilité causale » avec les observations moléculaires et neuro-cellulaires que les traits psychologiques dont la spécification fonctionnelle est exclusivement proximales. D'ailleurs, le fait d'avoir identifié un module anatomique plutôt qu'un simple module fonctionnel, permet d'envisager l'étude de marqueurs génétiques associés notamment aux structures cytoarchitecturales sous-jacentes à la rTPJ (Kanwisher, 2010). La découverte de ces hypothétiques marqueurs génétiques peut contribuer à l'identification de régions homologues chez les primates non-humains et ainsi ouvrir la porte à une comparaison de leurs fonctions. Toutefois, des données récentes suggèrent que la rTPJ est une région corticale ne possédant pas d'homologue simple chez les primates non-humains (Mars, Sallet, Neubert et Rushworth, 2013; Rushworth, Mars et Sallet, 2013).

En résumé, la notion historique d'adaptation implique plusieurs critères d'évidence ne

pouvant être satisfaits à la lumière des données scientifiques actuellement disponibles. En effet, seul le critère d'héritabilité semble être satisfait.

Dans la section suivante, nous adopterons une notion d'adaptation moins restrictive (i.e. la notion « ingénierique » d'adaptation) tout en éliminant l'hypothèse alternative qui propose l'apprentissage comme cause du développement ontogénétique d'une structure spécialisée pour l'éTdE.

2.4.2. Conception « ingénierique »

Selon Pinker et Bloom (1990), le fait qu'une structure exhibe un design fonctionnel complexe (*special design*) pour une fonction adaptative, ajouté au fait qu'aucun autre processus biologique ne puisse expliquer cette complexité indique clairement la présence d'une adaptation : « « Evolutionary theory offers clear criteria for when a trait should be attributed to natural selection: complex design for some function, and the absence of alternative processes capable of explaining such complexity » (p. 707). Selon Cosmides et Tooby (2005, p. 27), l'identification du design fonctionnel complexe d'un trait nécessite l'évaluation de la probabilité avec laquelle les propriétés de ce trait résolvent un problème adaptatif ancestral. Cependant, pour obtenir des évidences d'adéquation étroite (*tigh fit*) entre un trait et un problème adaptatif, il faut surmonter deux difficultés. Premièrement, l'adéquation entre certaines pressions sélectives ancestrales et certaines propriétés d'une structure peuvent être le résultat de différents scénarios causaux autres qu'une évolution par sélection naturelle (Andrews *et al.*, 2002). En effet, une structure peut avoir été initialement sélectionnée pour résoudre un problème adaptatif particulier, mais, grâce à des modifications écologiques ou à des capacités de construction de niche, exhibées une adéquation étroite avec les pressions sélectives d'un nouvel environnement sans que ces dernières soit la cause directe de cette adéquation. Par exemple, si une structure fut initialement sélectionnée pour être flexible au niveau ontogénétique, comme certains programmes

d'apprentissage évolués, cette structure peut être exaptée pour résoudre de nouveaux problèmes adaptatifs et tout de même exhibée, grâce à sa flexibilité, une adéquation avec ces nouvelles pressions sélectives. Deuxièmement, puisqu'une adaptation est toujours un équilibre (*trade-off*) entre plusieurs pressions sélectives et différentes contraintes génétiques et allométriques (Geary et Huffman, 2002), il est possible qu'une adaptation n'exhibe pas de manière probante une adéquation étroite avec les pressions sélectives ancestrales qui l'ont « façonnées ».

À la lumière de ces difficultés, Andrews *et al.* (2002) choisissent d'autres types d'évidences pour satisfaire le critère de design fonctionnel complexe (*special design*) (voir aussi Simpson et Campbell, 2005). Selon ces derniers, un trait doit plutôt effectuer *spécifiquement* et de manière *compétente* sa fonction pour être minimalement considéré comme un design fonctionnel complexe. Un avantage de cette conception est que selon cette formulation les adaptations n'ont pas à exhiber de signe clair de conception adéquate (*well designed*) pour effectuer de manière *compétente* leur effet, ce qui est cohérent avec le fait que les adaptations sont le produit de modifications d'anciennes adaptations (Barrett, 2012; Marcus, 2006).

Ayant déjà amplement discuté de la *spécificité* d'une sous-région de la rTPJ pour l'éTdE (section 2.3), si l'on veut établir que cette sous-région est une adaptation « ingénierique », il nous reste à vérifier si cette dernière participe de manière *compétente* à l'éTdE. Sur ce point, les données sont loin d'être convergentes. Très peu de données sont disponibles à propos des opérations (*workings*) algorithmiques de la rTPJ (Saxe, 2010; mais voir Koster-Hale et Saxe, 2013) et le peu de données dont nous disposons semblent contradictoires. D'un côté certains chercheurs modélisant les performances de la TdE, conformément à un modèle bayésien, concluent, à partir de leurs observations, que les inférences de la TdE « come surprisingly close to those of an ideal rational model. » (Baker, Saxe et Tenenbaum, 2011, p. 2473). De l'autre côté, on accumule plusieurs observations à propos de biais et d'erreurs d'attribution de la TdE (e.g. biais d'attribution fondamental, bais d'attribution égoïste) qui suggèrent que

les opérations de la TdE chez l'adulte sont loin d'atteindre des résultats optimaux (Apperly, 2011). Ces différents biais et erreurs d'attribution peuvent toutefois être réconciliés avec une interprétation évolutionniste si on les réinterprètent dans le cadre de la théorie de la gestion des erreurs (*error management theory*) (voir Haselton et Buss, 2000). L'idée principale de cette théorie est que la sélection naturelle favorisera l'évolution des règles d'inférences les plus bénéfiques ou les moins coûteuses, même si ces règles d'inférences sont biaisées et commettent fréquemment des erreurs. Par exemple, le cas du biais de surestimation de l'intérêt sexuel (*sexual overperception*) par les hommes, où les hommes surestime systématiquement l'intérêt sexuel des femmes à leur égard, est interprété, dans le cadre de la théorie de la gestion des erreurs, comme un biais ultimement bénéfiques.

D'autres données supportant la *compétence* de la structure spécialisée pour l'éTdE, peuvent découler du retour à des mesures de l'exactitude des performances psychologiques (voir Zaki et Ochsner, 2011). En effet, la mesure des taux d'exactitude et de la validité des prédictions de l'éTdE peuvent nous procurer des indications sur les performances écologiques de l'éTdE en contexte intra-groupe et inter-groupe. Ces informations pourraient ensuite être intégrées aux différentes hypothèses concernant l'EAE de la TdE (e.g., l'hypothèse de l'intelligence sociale, l'hypothèse de l'intelligence socio-écologique, l'hypothèse de l'intelligence machiavélienne) ce qui pourrait aider à déterminer si l'éTdE effectue de manière *compétente* sa fonction.

2.4.2.1. L'apprentissage comme hypothèse alternative

Un aspect important de notre stratégie argumentative doit être souligné : jusqu'à maintenant, nous avons examiné si les données scientifiques concernant une structure spécialisée pour l'éTdE satisfont une liste de critères préspécifiés qui permettent de justifier l'attribution du statut d'adaptation. Toutefois, selon Williams (1966), pour

déterminer la validité d'une hypothèse adaptative, il serait préférable de démontrer l'implausibilité de l'ensemble des hypothèses alternatives expliquant le parcours phylogénétique d'un trait. En ce qui concerne notre hypothèse en faveur du statut d'adaptation de la sous-région de la rTPJ, il faudrait donc pouvoir éliminer les hypothèses alternatives de sa spécialisation fonctionnelle proximale. D'ailleurs, l'accomplissement *spécifique* et *compétent* d'une fonction par une structure peut être le résultat de scénario causaux autres que la sélection naturelle. Une capacité générique d'apprentissage (*domain-general learning*) est le candidat par excellence d'une hypothèse alternative voulant expliquer comment une structure peut accomplir, de manière *spécifique* et *compétente*, un effet sans avoir été favorisé par la sélection naturelle. Par exemple, Cosmides et Tooby (2005), dans leur défense des adaptations neurocognitives, soutiennent : « If computational specializations for social exchange are acquired via some general-purpose learning process, then we should not claim that the specialization is an evolded adaptation for social exchange. Instead, the social exchange specialization would be the product of a learning mechanism that evolved to solve a different, perhaps more general, adaptive problem » (p. 616). Afin d'éviter cet écueil, Cosmides et Tooby (2005) mettent l'emphase sur le fait que les structures psychologiques sélectionnées doivent performer des inférences basées sur une logique évolutionniste plutôt que sur une logique formelle apprise à partir de l'expérience. Précisément, l'argument est que les patterns de performance et d'erreurs produits par les algorithmes darwiniens ne peuvent pas découler de mécanismes d'apprentissage inductif généraux (e.g., bayésien, régression multiple, calcul de contingence) s'informant de la distribution statistique (expérience, répétition, familiarité) de l'information offerte par l'environnement ontogénétique. En effet, ces patterns de performance et d'erreur doivent être inférables grâce à des analyses de stratégies évolutivement stables (SES) appliquées à l'EAE des *Homo sapiens*. Autrement dit, les algorithmes d'un module darwinien sont inférables à partir de la distribution statistique de l'information présente dans l'EAE et non à partir de celle de l'environnement ontogénétique actuel. Cependant, cette détermination génétique des

caractéristiques computationnelles n'implique qu'une seule contrainte développementale, soit que ce développement soit fiable. Ce développement fiable peut même être majoritairement influencé par les interactions environnementales, il suffit que ces propriétés environnementales aient été stables durant l'évolution phylogénétique pour que les programmes développementaux soient dépendants, pour leurs développements normaux, à ces propriétés environnementales stables. Il est important de remarquer qu'il est probable que les processus ontogénétiques permettant le développement de l'organisation corticale spécifique aux différentes espèces découlent, du moins en partie, d'une sélection pour une niche sensorielle particulière (e.g. nocturne ou diurne) (Barton, 2007; voir aussi Aboitiz et Zamorano, 2013).

Dans cette sous-section, nous examinerons des données interculturelles du développement ontogénétique de l'éTdE afin de renforcer notre position en faveur de l'existence d'une adaptation « ingénierique » spécialisée pour l'éTdE, tout en discréditant la possibilité qu'une capacité générique d'apprentissage soit à la base de cette spécialisation fonctionnelle proximale. En effet, sachant que l'environnement physique et social dans lequel les enfants se développent varie entre les cultures et qu'une capacité générique d'apprentissage est sensible aux conditions environnementales (Machery, à venir), si le développement d'une structure spécialisée pour l'éTdE était le résultat d'une capacité générique d'apprentissage plutôt que le produit de la sélection naturelle, alors on devrait pouvoir observer des variations dans le développement de cette structure en fonction des différentes conditions environnementales dans lesquels l'éTdE se développe. Au contraire, certains aspects du développement de la rTPJ suggèrent que son développement est relativement indépendant des informations environnementales. Par exemple, Bedny, Pascual-Leone et Saxe (2009) observe que la rTPJ se développe normalement (i.e. préservation de localisation et du profil d'activation neurologique) chez des aveugles congénitaux. Cependant, les enfants sourds, n'apprenant le langage des signes que très

tard, possèdent un retard significatif de l'âge de réussite des tâches de fausses croyances (Woolfe, Want et Siegal, 2002). Les facteurs culturels ne semblent pas influencer l'âge de la réussite des tâches de fausses croyances. Par exemple, Callaghan, Rochat, Lillard, Claux, Odden, *et al.* (2005), après avoir testés sur cinq cultures différentes (Pérou, Inde, Samoa, Thaïlande, Canada) la même tâche de fausses croyances et avoir observés une synchronie similaire entre ces 5 cultures (i.e., la réussite de la tâche entre l'âge de 3 et 5 ans) concluent à l'identification d'un jalon cognitif universel dans le développement des capacités de TdE. Cependant, certaines études mesurent des taux d'échec à la tâche des fausses croyances allant jusqu'à 60% pour des enfants de 5-6 ans de culture japonaise (voir Kobayashi et Temple, 2009). De plus, une méta-analyse comparant les enfants chinois et les enfants nord-américains démontre une grande variation dans le délai de réaction associé aux performances à des tâches de fausses croyances (Liu, Wellman, Tardif et Sabbagh, 2008). Ces différents résultats suggèrent que certaines étapes développementales centrales de la TdE serait indépendantes de l'expérience et que d'autres aspects seraient flexibles et ouverts aux influences culturelles.

En ce qui concerne la rTPJ, Kobayashi et Temple (2009) remarquent une diminution significative de l'activation des TPJ chez les adultes japonais par rapport aux adultes américains. Selon Kobayashi et Temple (2009), cette différence interculturelle d'activation du TPJ chez les adultes s'expliquerait par une différence culturelle de l'approche de la TdE. Précisément, de manière compatible avec les données de Callaghan *et al.* (2005), il semble que la maturation de la TdE chez l'adulte soit sensible à certains facteurs culturels et se raffine en fonction de ces différents facteurs. Ces observations d'une maturation de certains aspects centraux la TdE indépendamment des variations culturelles et d'une influence culturelle sur les raffinements plus tardifs de la TdE sont compatibles avec les observations concernant les variations culturelles de l'activation des TPJ. En effet, les TPJ continuent à se développer (i.e. à modifier son profil d'activation en se spécialisant de plus en plus au

domaine des métareprésentations épistémiques) jusqu'à très tard dans l'adolescence (Saxe, 2010). De plus, la trajectoire développementale typique des performances aux tâches des fausses croyances semble correspondre à la progressive spécialisation de la rTPJ pour certaines types d'états mentaux spécifiques. Par exemple, chez les adultes, la rTPJ limite son activation aux métareprésentations mentales, tandis que, chez les enfants de 6 à 9 ans, elle s'active aussi pour les récits incluant des personnes (Saxe, 2010).

En définitive, nous retrouvons une polarisation de la question des adaptations psychologiques entre certains auteurs qui vont considérer que les attributions du statut d'adaptation à des traits phénotypiques psychologiques doivent donc être mis en suspens tant que plus de données empiriques ne seront pas recueillies (Lloyd, 1999)[40], d'autres (Carruthers, 2006, p16) qui vont rejeter du revers de la main ce scepticisme en reprenant l'argument selon lequel l'évolution par sélection naturelle demeure l'unique explication valable d'une telle complexité fonctionnelle organisée (voir Atran, 2005 pour une critique de cet argument), et d'autres (Griffiths, 2007) qui vont proposer de mettre l'emphase sur les traits homologues afin de maximiser nos chances de pouvoir ultimement justifier nos ambitions de découvrir des adaptations neurocognitives chez l'*Homo sapiens*. Nous pensons contribuer au projet de Griffiths en maintenant, malgré le peu de données pour la corroborer, notre hypothèse que la sous-région de la rTPJ spécialisée pour l'éTdE est une structure ayant été « façonnée » par la sélection naturelle. Puisque la rTPJ possède sa propre trajectoire développementale et ses propres facteurs d'activation, il est plausible qu'elle représente un module développemental possédant sa propre histoire évolutive et constituant une unité de sélection. De manière plus générale, il est plausible que plusieurs modules darwiniens soient spécifiques à des types distincts de contexte

[40] « It is *not scientifically acceptable* within evolutionary biology to conclude that, because a given pattern of responses contributes to evolutionary success, then there is some 'organ' (or part of the brain) producing such a pattern, that is therefore an adaptation (see Williams, 1966). This is because the 'organ' or 'module' may not actually exist as a biologically real trait, and even if it does, its current function may or may not be the same as the past function(s) » (Lloyd, 1999, p. 224).

interactionnel avec différentes entités intentionnelles, tels que les contextes sociaux d'interaction intra-espèces ou les contextes de prédation inter-espèces. Ce qui rend complémentaire plusieurs scénarios de l'origine phylogénétique de la TdE puisque cette dernière serait finalement constituée de plusieurs adaptations psychologiques différentes se coordonnant.

En résumé, dans ce chapitre, nous avons caractérisé l'attribution d'états mentaux épistémiques (section 2.2). Nous avons établi l'existence, à l'aide d'un ensemble de données convergentes, d'une région corticale spécialisée pour cette cette capacité d'attribution (section 2.3). Nous avons finalement évalué la plausibilité de l'hypothèse selon laquelle cette structure spécialisée serait une adaptation historique (section 2.4.1) ou encore une adaptation « ingénierique » (section 2.4.2). Le tout dans le but d'établir que l'éTdE est bel et bien sous-tendue par un module darwinien (voir section 1.2). À la lumière des données scientifiques actuellement disponibles, il ne nous est pas possible de conclure que la structure spécialisée pour l'éTdE est une adaptation historique. En revanche, le statut moins restrictif d'adaptation « ingénierique » est un candidat plausible. En effet, l'éTdE est effectuée de manière *spécifique* et *compétente* par une sous-région de la rTPJ et il est peu probable que cette *spécificité* et cette *compétence* découle d'une capacité générique d'apprentissage considérant le parcours développemental de la spécialisation pour l'éTdE.

CONCLUSION

Dans ce mémoire, nous avons distingué différents usages de la notion de modularité ayant cours en sciences cognitives. Ces distinctions nous ont permis de caractériser la notion de module darwinien comme une structure psychologique spécialisée pour une fonction et ayant été sélectionné par la sélection naturelle pour effectuer cette spécialisation fonctionnelle. De plus, nous avons vu que les propriétés attendues des modules darwiniens dépendent, à la fois, de leur fonction adaptative particulière (i.e les contraintes téléologiques) et de leur histoire évolutive particulière (i.e. les contraintes phylogénétiques). En effet, nous avons établi que la modularité darwinienne, malgré l'abandon des propriétés fodoriennes de la modularité cognitive, ne se réduit pas à la notion moins contentieuse de module fonctionnel ou anatomique. À la lumière de la nature distincte de la notion de module darwinien, nous avons ensuite tenté d'établir la pertinence heuristique de cette notion pour la décomposition de la cognition humaine en démontrant qu'il est plausible qu'une sous-capacité de la cognition sociale humaine (i.e. l'attribution d'états mentaux épistémiques (éTdE)) soit effectivement sous-tendue par un module darwinien. Ce travail de recherche permet d'établir qu'il est plausible qu'une région corticale spécialisée dans l'attribution d'états mentaux soit « façonnée » par la sélection naturelle.

Nous identifions deux limites de notre position. Premièrement, en raison des difficultés associées à l'identification des pressions sélectives exactes ayant façonnées la structure spécialisée pour l'éTdE, notre caractérisation de l'éTdE se limite à une caractérisation proximale des aspects ontogénétiques et mécanistiques. En effet, rien ne garantit que cette caractérisation proximale corresponde effectivement à la capacité évoluée qui fut sélectionnée dans les populations ancestrales (e.g. voir Neuberg, Kenrick et Shaller, 2010 pour une caractérisation évolutionniste de la

cognition sociale humaine). Deuxièmement, nous sommes conscients que même s'il s'avérait exact que l'éTdE soit sous-tendue par un module darwinien, cela ne confirmerait aucunement la véracité de l'aHMM. En effet, une seule instance ne suffit pas à soutenir une hypothèse postulant l'existence d'une multitude de modules darwiniens au centre de l'esprit. Nous croyons toutefois que cette éventualité militerait fortement en faveur de, non seulement, la plausibilité de l'aHMM, mais aussi la valeur heuristique de la modularité darwinienne pour la décomposition de la cognition humaine.

BIBLIOGRAPHIE

Abell, F., Krams, M., Ashburner, J., Passingham, R., Friston, K., Frackowiak, R., Happé, F., Frith, C. et Frith, U. (1999). The neuroanatomy of autism : a voxel-based whole brain analysis of structural scans. *Neuroreport*, 10(8), 1647-1651.

Aboitiz, F., Aboitiz, S. et García, R. (2010). The Phonological Loop: A Key Innovation in Human Evolution. *Current Anthropology*, 51(S1), S55-S65.

Aboitiz, F., et Zamorano, F. (2013). Neural progenitors, patterning and ecology in neocortical origins. *Frontiers in neuroanatomy*, 7(38), 1-15.

Adams, M.P. (2011). Modularity, Theory of Mind, and Autism Spectrum Disorder. *Philosophy of Science*, 78(5), 763-773.

Adams, M.P. (2013). Explaining the theory of mind deficit in autism spectrum disorder. *Philosophical Studies*, 163(1), 233-249.

Amodio, D.M. et Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268-277.

Anderson, M.L. (2006). Evidence for massive redeployment of brain areas in cognitive function. *Proceedings of the Cognitive Science Society*, Actes du colloque 2006, (Vol.28), 24-29.

Anderson, M.L. (2007). The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology*, 20(2), 143-174.

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

Andrews, P.W., Gangestad, S.W. et Matthews, D. (2002). Adaptationism-how to carry out an exaptationist program. *Behavioral and Brain Sciences*, 25(4), 489-504.

Apperly, I.A. (2008). Beyond simulation–theory and theory–theory: why social cognitive neuroscience should use its own concepts to study "Theory of Mind". *Cognition*, 107(1), 266-283.

Apperly, I.A. (2010). *Mindreaders: The cognitive basis of "theory of mind"*. Hove, UK : Psychology Press.

Apperly, I.A. et Butterfill, S.A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4), 953.

Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C. et Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, 17(10), 841-844.

Apperly, I.A., Samson, D., Chiavarino, C., Bickerton, W.-L. et Humphreys, G.W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal,"reality-unknown" false belief and false photograph tasks. *Cognition*, 103(2), 300-321.

Apperly, I.A., Samson, D., Chiavarino, C. et Humphreys, G.W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, 16(10), 1773-1784.

Apperly, I.A., Samson, D. et Humphreys, G.W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology*, 45(1), 190.

Ariew, A. (2003). Ernst Mayr's' ultimate/proximate'distinction reconsidered and reconstructed. *Biology and Philosophy*, 18(4), 553-565.

Atique, B., Erb, M., Gharabaghi, A., Grodd, W. et Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing. *Neuroimage*, 55(4), 1899-1911.

Atkinson, A.P. et Adolphs, R. (2011). The neuropsychology of face perception: beyond simple dissociations and functional selectivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571), 1726-1738.

Atkinson, A.P. et Wheeler, M. (2004). The Grain of Domains : The Evolutionary-Psychological Case Against Domain-General Cognition. *Mind & Language*, 19(2), 147-176.

Atran, S. (2005). Adaptationism for human cognition: Strong, spurious or weak? *Mind & Language*, 20(1), 39-67.

Baars, B.J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baker, C.L., Saxe, R.R. et Tenenbaum, J.B. (2011). Bayesian theory of mind : Modeling joint belief-desire attribution. *Proceedings of the thirty-second annual conference of the cognitive science society*, Actes du colloque 2011, 2469-2474.

Barkow, J.H., Cosmides, L.E. et Tooby, J.E. (1992). T*he adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge : MIT Press.

Baron-Cohen, S. (2009). Autism : The Empathizing–Systemizing (E-S) Theory. *Annals of the New York Academy of Sciences*, 1156(1), 68-80.

Baron-Cohen, S., Leslie, A.M. et Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1), 37-46.

Baron-Cohen, S., Leslie, A.M. et Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, 4(2), 113-125.

Baron-Cohen, S., Wheelwright, S., Stone, V. et Rutherford, M. (1999). A mathematician, a physicist and a computer scientist with Asperger syndrome : Performance on folk psychology and folk physics tests. *Neurocase*, 5(6), 475-483.

Barrett, H.C. (2005). Enzymatic computation and cognitive modularity. *Mind & Language*, 20(3), 259-287.

Barrett, H.C. (2006). Modularity and design reincarnation. Dans P. Carruthers, S. Laurence et S. Stich (dir.), *The innate mind : volume 2, culture and cognition* (Vol. 2, p. 199-217). New York : Oxford University Press.

Barrett, H.C. (2007). Development as the target of evolution: A computational approach to developmental systems. Dans S. Gangestad et J. Simpson (dir.), *The evolution of mind: Fundamental questions and controversies* (p. 186-192). Guilford Press.

Barrett, H.C. (2009). Where there is an adaptation, there is a domain: the form-function fit in information processing. Dans S.M. Platek et T. K. Shackleford (dir.), *Foundations in Evolutionary Cognitive Neuroscience* (p. 97-116). Cambridge : Cambridge University Press.

Barrett, H.C. (2012). A hierarchical model of the evolution of human brain specializations. *Proceedings of the national Academy of Sciences*, 109(Supplement 1), 10733-10740.

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

Barrett, H.C., Todd, P.M., Miller, G.F. et Blythe, P.W. (2005). Accurate judgments of intention from motion cues alone : A cross-cultural study. *Evolution and Human Behavior*, 26(4), 313-331.

Barrett, L., Henzi, P. et Rendall, D. (2007). Social brains, simple minds: does social complexity really require cognitive complexity? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 561-575.

Barton, R. (2007). Evolutionary specialization in mammalian cortical structure. *Journal of evolutionary biology*, 20(4), 1504-1511.

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

Bechtel, W. (2002). Decomposing the mind-brain : A long-term pursuit. *Brain and Mind*, 3(2), 229-242.

Bechtel, W. (2003). Modules, brain parts, and evolutionary psychology. Dans S. J. Scher et F. Rauscher (dir.), *Evolutionary psychology : Alternative approaches* (p. 211-227). Dordrecht, the Netherlands : Kluwer.

Bechtel, W. et Richardson, R.C. (1993). *Discovering complexity : Decomposition and localization as scientific research strategies.* Princeton, New Jersey: Princeton University Press.

Bedny, M., Pascual-Leone, A. et Saxe, R.R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences*, 106(27), 11312-11317.

Behme, C. et Deacon, S.H. (2008). Language Learning in Infancy: Does the Empirical Evidence Support a Domain Specific Language Acquisition Device? *Philosophical Psychology*, 21(5), 641-671.

Bergeron, V. (2007). Anatomical and functional modularity in cognitive science : Shifting the focus. *Philosophical Psychology*, 20(2), 175-195.

Bergeron, V. (2008). *Cognitive architecture and the brain: Beyond domain-specific functional specification.* (Thèse de doctorat). University of British Columbia. Récupéré de CIRCLE, l'archive de publications électroniques de UBC. https://circle.ubc.ca/handle/2429/2711?show=full

Bloom, P. et German, T.P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1), B25-B31.

Bock, W.J. (1980). The definition and recognition of biological adaptation. *American Zoologist*, 20(1), 217-227.

Bolhuis, J.J., Brown, G.R., Richardson, R.C. et Laland, K.N. (2011). Darwin in mind: new opportunities for evolutionary psychology. *PLoS biology*, 9(7), e1001109.

Bowler, D.M. et Thommen, E. (2000). Attribution of mechanical and social causality to animated displays by children with autism. *Autism*, 4(2), 147-171.

Boyer, P. (2000). Natural epistemology or evolved metaphysics? Developmental evidence for early-developed, intuitive, category-specific, incomplete, and stubborn metaphysical presumptions. *Philosophical psychology*, 13(3), 277-297.

Boyer, P. et Barrett, H.C. (2005). Evolved intuitive ontology: Integrating neural, behavioral and developmental aspects of domain-specificity. Dans D. Buss (dir.), *The handbook of evolutionary psychology* (p. 96-118). New Jersey : Wiley.

Brandon, R.N. (1990). *Adaptation and environment*. Princeton, New Jersey : Princeton University Press.

Brandon, R.N. (1999). The units of selection revisited: the modules of selection. *Biology and Philosophy*, 14(2), 167-180.

Brandon, R.N. (2005). Evolutionary modules : Conceptual analyses and empirical hypotheses. Dans Callebaut, W. et D. Rasskin-Gutman (dir.), *Modularity : understanding the development and evolution of natural complex systems*. (p. 51-60). Cambridge, MA : MIT Press.

Buckner, R.L. et Krienen, F.M. (2013). The evolution of distributed association networks in the human brain. *Trends in cognitive sciences*, 17(12), 648-665.

Buller, D.J. (1993). Confirmation and the computational paradigm (or : Why do you think they call it artificial intelligence?). *Minds and Machines*, 3(2), 155-181.

Buller, D.J. (2005). *Adapting minds : Evolutionary psychology and the persistent quest for human nature*. MA : MIT Press.

Buller, D.J. et Hardcastle, V. (2000). Evolutionary psychology, meet developmental neurobiology : Against promiscuous modularity. *Brain and Mind*, 1(3), 307-325.

Buss, D.M. (1989). Sex differences in human mate preferences : Evolutionary hypotheses tested in 37 cultures. *Behavioral and brain sciences*, 12(1), 1-49.

Buss, D.M., Haselton, M.G., Shackelford, T.K., Bleske, A.L. et Wakefield, J.C. (1998). Adaptations, exaptations, and spandrels. *American psychologist*, 53(5), 533.

Byrne, R. et Whiten, A. (1989). *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*, Clarendon Press.

Calabretta, R. et Parisi, D. (2005). Evolutionary Connectionism and Mind/Brain Modularity. Dans W. Callebaut et D. Rasskin-Gutman (dir.), *Modularity :*

Understanding the development and evolution of complex natural systems (p. 309–330.). Cambridge, MA : MIT Press.

Call, J. et Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in cognitive sciences*, 12(5), 187-192.

Callaghan, T., Rochat, P., Lillard, A., Claux, M.L., Odden, H., Itakura, S., Tapanya, S. et Singh, S. (2005). Synchrony in the onset of mental-state reasoning evidence from five cultures. *Psychological Science*, 16(5), 378-384.

Callebaut, W.G. et Rasskin-Gutman, D. (2005). *Modularity: understanding the development and evolution of natural complex systems*. Cambridge, MA : MIT Press.

Carrington, S.J. et Bailey, A.J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human brain mapping*, 30(8), 2313-2335.

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

Castelli, F., Frith, C., Happé, F. et Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125(8), 1839-1849.

Chater, N. (2003). How much can we learn from double dissociations? *Cortex*, 39(1), 167-169.

Cheng, P.W. et Holyoak, K.J. (1989). On the natural selection of reasoning theories. *Cognition*, 33(3), 285-313.

Chiappe, D. et MacDonald, K. (2005). The evolution of domain-general mechanisms in intelligence and learning. *The Journal of general psychology*, 132(1), 5-40.

Chomsky, N. (1959). A review of B.F. Skinner's Verbal Behavior. *Language*, 35(1), 26-58.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, 3(1), 1-15.

Chomsky, N. (1984). *Modular Approaches to the Study of the Mind*. San Diego : San Diego State University.

Clark, A. (1999). An embodied cognitive science? *Trends in cognitive sciences*, 3(9), 345-351.

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

431.

Clavagnier, S., Falchier, A. et Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. Cognitive, Affective, & Behavioral Neuroscience, 4(2), 117-126.

Cohen, L. et Dehaene, S. (2004). Specialization within the ventral stream: the case for the visual word form area. *Neuroimage*, 22(1), 466-476.

Coltheart, M. (1999). Modularity and cognition. *Trends in cognitive sciences*, 3(3), 115-120.

Coltheart, M. (2001). Assumptions and methods in cognitive neuropsychology. Dans Rapp, B. (dir.), *The handbook of cognitive neuropsychology* (p. 3-21). Hove, UK : Psychology Press.

Coltheart, M. (2011). Methods for modular modelling: Additive factors and cognitive neuropsychology. *Cognitive neuropsychology*, 28(3-4), 224-240.

Corbetta, M., Patel, G. et Shulman, G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3), 306-324.

Cosmides, L. (1985). *Deduction or Darwinian Algorithms*? Harvard University, Cambridge, Massachusetts. http://www.cep.ucsb.edu/papers/cosmides_1985_intro.pdf.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3), 187-276.

Cosmides, L., Barrett, H.C. et Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 9007-9014.

Cosmides, L. et Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. Dans J. Dupre (dir.), *The latest and best: Essays on evolution and optimality* (p. 277-306). Cambridge, MA : MIT Press.

Cosmides, L. et Tooby, J. (1989). Evolutionary psychology and the generation of culture, part II : Case study : A computational theory of social exchange. *Ethology and sociobiology*, 10(1), 51-97.

Cosmides, L. et Tooby, J. (1992). Cognitive adaptations for social exchange. Dans J.H. Barkow, L.E. Cosmides et J.E. Tooby (dir.), T*he adapted mind : Evolutionary psychology and the generation of culture* (p. 163-228). Oxford

University Press.

Cosmides, L. et Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. Dans L.A. Hirschfeld et S.A. Gelman (dir.), *Mapping the mind: Domain specificity in cognition and culture* (p. 85-116). Cambridge University Press.

Cosmides, L. et Tooby, J. (1995). From function to structure: The role of evolutionary biology and computational theories in cognitive neuroscience. Dans M. Gazzaniga (dir.), *The cognitive neurosciences* (p. 1199-1210). Cambridge, MA : MIT Press.

Cosmides, L. et Tooby, J. (2000). Consider the source: The evolution of adaptations for decoupling and metarepresentation. Dans D. Sperber (dir.), *Metarepresentations: A multidisciplinary perspective* (p. 53-115). Oxford University Press.

Cosmides, L. et Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. Dans D. Buss (dir.), *The handbook of evolutionary psychology* (p. 584-627). New Jersey : Wiley.

Cummins, R. (2000). How does it work?" versus" what are the laws?": Two conceptions of psychological explanation. Dans F.C. Keil et R.A. Wilson (dir.), *Explanation and cognition* (p. 117-144). MIT Press.

Currie, G. et Sterelny, K. (2000). How to think about the modularity of mind-reading. *The Philosophical Quarterly*, 50(199), 145-160.

Daly, M. et Wilson, M.I. (1999). Human evolutionary psychology and animal behaviour. *Animal Behaviour*, 57(3), 509-519.

Darwin, C. R. (1845). *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N.* 2nd edition. London : John Murray.

de Beeck, H.P.O., Haushofer, J. et Kanwisher, N.G. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Reviews Neuroscience*, 9(2), 123-135.

Decety, J. et Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13(6), 580-593.

Dehaene, S. et Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56(2), 384-398.

Dehaene, S. et Cohen, L. (2011). The unique role of the visual word form area in

reading. *Trends in cognitive sciences*, 15(6), 254-262.

Dehaene, S. et Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1), 1-37.

Dennett, D.C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4), 87-106.

Dennett, D.C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1(04), 568-570.

Dennett, D.C. (1987). *The Intentional Stance*. Cambridge, MA : MIT Press.

Dobson, S.D. et Sherwood, C.C. (2011). Correlated evolution of brain regions involved in producing and processing facial expressions in anthropoid primates. *Biology letters*, 7(1), 86-88.

Downes, S.M. (2005). Integrating the multiple biological causes of human behavior. *Biology and Philosophy*, 20(1), 177-190.

Duchaine, B., Cosmides, L. et Tooby, J. (2001). Evolutionary psychology and the brain. *Current opinion in neurobiology*, 11(2), 225-230.

Duchaine, B.C., Yovel, G., Butterworth, E.J. et Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms : Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, 23(5), 714-747.

Dunbar, R.I.M. et Barrett, L. (2007). Evolutionary psychology in the round. Dans R.I.M. Dunbar et L. Barrett (dir.), *Oxford handbook of evolutionary psychology* (p. 3-9). Oxford University Press.

Dunbar, R.I.M. et Shultz, S. (2007). Evolution in the social brain. *Science*, 317(5843), 1344-1347.

Duntley, J.D. (2005). Adaptations to dangers from humans. Dans D. Buss (dir.), T*he handbook of evolutionary psychology* (p. 224-249). New Jersey : Wiley.

Eastwick, P.W. (2009). Beyond the pleistocene: using phylogeny and constraint to inform the evolutionary psychology of human mating. *Psychological bulletin*, 135(5), 794.

Egeth, M. et Kurzban, R. (2009). Representing metarepresentations: Is there Theory of Mind-specific cognition? *Consciousness and cognition*, 18(1), 244-254.

Elman, J., Bates, E.J., Johnson, M. M., Karmiloff-Smith, A., Parisi, D. et Plunkett, K. (1996). *Rethinking innateness : a connectionist perspective on development.*

Cambridge, MA : MIT Presss.

Emery, N.J. et Clayton, N.S. (2009). Comparative social cognition. *Annual review of psychology*, 60, 87-113.

Falchier, A., Clavagnier, S., Barone, P. et Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience*, 22(13), 5749-5759.

Farah, M.J. (1994). Neuropsychological inference with an interactive brain : A critique of the" locality" assumption. *Behavioral and Brain Sciences*, 17(1), 43-60.

Faucher, L. et Poirier, P. (2009). Modularité et psychologie évolutionniste. Dans J.-B. Van der Henst et H. Mercier (dir.), *Darwin en tête! L'évolution et les sciences cognitives* (p. 275-308). Grenoble : Presses Universitaires de Grenoble.

Faucher, L. et Tappolet, C. (2006). Introduction : Modularity and the nature of emotions. *Canadian Journal of Philosophy*, 36(5), vii-xxxi.

Fessler, D.M. et Machery, E. (2012). Culture and cognition. Dans E. Margolis, R. Samuels et S. Stich (dir.), *The Oxford Handbook of Philosophy of Cognitive Science* (p. 503-527). New York : Oxford University Press.

Fisch, G.S. (2013). Autism and epistemology IV : Does autism need a theory of mind? *American Journal of Medical Genetics Part A*, 161(10), 2464-2480.

Fodor, J.A. (2005). Reply to Steven Pinker 'So how does the mind work?'. *Mind & Language*, 20(1), 25-32.

Fodor, J.A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA : MIT Press.

Fodor, J.A. (2001). *The mind doesn't work that way: the scope and limits of computational psychology*. Cambridge, MA : MIT press.

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

Frankenhuis, W.E. et Ploeger, A. (2007). Evolutionary psychology versus Fodor : Arguments for and against the massive modularity hypothesis. *Philosophical Psychology*, 20(6), 687-710.

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

Frith, C.D. (2004). Schizophrenia and theory of mind. *Psychological medicine*, 34(03), 385-389.

Gall, F.J. (1822–1825). *Sur les fonctions du cerveau et sur celles de chacune de ses parties, avec des observations sur la possibilité de reconnaître les instincts, les penchants, les talents ou les dispositions morales et intellectuelles des hommes et des animaux, par la configuration de leur cerveau et de leur tête*. Paris : J.B. Ballière.

Gallese, V. et Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12), 493-501.

Gallese, V., Keysers, C. et Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in cognitive sciences*, 8(9), 396-403.

Gallese, V. et Rochat, M. (2010). Motor Cognition. Dans D.Z. Zelazo, M. Chandler et E. Crone (dir.), *Developmental Social Cognitive Neuroscience*. The Jean Piaget Symposium Series (p. 13-41). New York : Psychology Press.

Gallistel, C.R. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. Dans M.S. Gazzaniga (dir.) *The new cognitive neurosciences* (p. 1179-1191). Cambridge, MA : MIT Press.

García, C.L. (2010). Functional Homology and Functional Variation in Evolutionary Cognitive Science. *Biological Theory*, 5 (2), 124-135.

Gaulin, S.J., Bock, G. et Cardew, G. (1997). Cross-cultural patterns and the search for evolved psychological mechanisms. Dans G. R. Bock et G. Cardew (dir.), *Characterizing human psychological adaptations*, Ciba Foundation Symposium 208 (p. 195-207). Chichester, England : Wiley.

Geary, D.C. et Huffman, K.J. (2002). Brain and cognitive evolution: forms of modularity and functions of mind. *Psychological bulletin*, 128(5), 667.

Gerrans, P. (2002). Modularity reconsidered. *Language & Communication*, 22(3), 259-268.

Gerrans, P. (2002). The theory of mind module in evolutionary psychology. *Biology and Philosophy*, 17(3), 305-321.

Gerrans, P. et Stone, V.E. (2008). Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *The British Journal for the Philosophy of Science*, 59(2), 121-141.

Gigerenzer, G. (1997). The modularity of social intelligence. Dans A. Whiten et R.W. Byrne (dir.), *Machiavellian intelligence II: extensions and evaluations* (p. 264-

288). Cambridge, UK: Cambridge University Press.

Gigerenzer, G. et Goldstein, D.G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4), 650.

Gigerenzer, G. et Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, 43(2), 127-171.

GilWhite, F. (2001). Are ethnic groups biological "species" to the human brain? *Current anthropology*, 42(4), 515-553.

Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J. et Haxby, J.V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, 19(11), 1803-1814.

Goldman, A.I. (1989). Interpretation Psychologized. Mind & Language, 4(3), 161-185.

Goldman, A.I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford : Oxford University Press.

Goldstein, D.G. et Gigerenzer, G. (2002). Models of ecological rationality : the recognition heuristic. *Psychological review*, 109(1), 75.

Gopnik, A. et Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in cognitive sciences*, 8(8), 371-377.

Gordon, R.M. (1986). Folk psychology as simulation. *Mind & Language*, 1(2), 158-171.

Gordon, R.M. (1995). Simulation without introspection or inference from me to you. Dans M. Davies and T. Stone (dir), *Mental simulation : Evaluations and Applications* (p. 53-67). Oxford : Blackwell.

Gordon, R.M. (2008). Beyond mindreading. *Philosophical explorations*, 11(3), 219-222.

Gould, S.J. et Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm : a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161), 581-598.

Gould, S.J. et Vrba, E.S. (1982). Exaptation : a missing term in the science of form. *Paleobiology*, 8, 4-15.

Gray, R., Heaney, M. et Fairhall, S. (2003). Evolutionary Psychology and the

challenge of adaptive explanation. Dans K. Sterelny et J. Fitness (dir.), *From Mating to Mentality* (p. 247–268). Psychology Press.

Griffith, E.M., Pennington, B.F., Wehner, E.A. et Rogers, S.J. (1999). Executive functions in young children with autism. *Child development*, 70(4), 817-832.

Griffiths, P.E. (1993). Functional analysis and proper functions. *The British Journal for the Philosophy of Science*, 44(3), 409-422.

Griffiths, P.E. (1996). The historical turn in the study of adaptation. *The British journal for the philosophy of science*, 47(4), 511-532.

Griffiths, P.E. (2006). Function, Homology, and Character Individuation. *Philosophy of Science*, 73(1), 1-25.

Griffiths, P.E. (2007). Evo-Devo Meets the Mind: Toward a Developmental Evolutionary Psychology. Dans R. Sanson et R.N. Brandon (dir.), *Integrating Development and Evolution* (p. 195-225) Cambridge, UK : Cambridge University Press.

Hagen, E.H. (2005). Controversial issues in evolutionary psychology. Dans D. Buss (dir.), *The handbook of evolutionary psychology* (p. 145-173). New Jersey : Wiley.

Happé, F., Ronald, A. et Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature neuroscience*, 9(10), 1218-1220.

Hardcastle, V.G. et Stewart, C.M. (2002). What do brain data really show? *Philosophy of Science*, 69(S3), S72-S82.

Hartline, H.K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology*. 121, 400-415.

Haselton, M.G. et Buss, D.M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology*, 78(1), 81.

Haselton, M.G. et Funder, D.C. (2006). The evolution of accuracy and bias in social judgment. Dans M. Schaller, D.T. Kenrick, et J.A. Simpson (dir.), *Evolution and social psychology* (p. 15-37). New York : Psychology Press.

He, Z., Bolz, M. et Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental science*, 14(2), 292-305.

Heider, F. et Simmel, M. (1944). An experimental study of apparent behavior. *The*

American Journal of Psychology, 57(2), 243-259.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in cognitive sciences*, 10(2), 64-69.

Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B. et Tomasello, M. (2007). Humans have evolved specialized skills of social cognition : the cultural intelligence hypothesis. *Science*, 317(5843), 1360-1366.

Hirschfeld, L. et Gelman, S. (1994) Toward a topography of mind : an introduction to domain specificity. Dans L.A Hirschfeld et S.A. Gelman (dir.), *Mapping the mind : Domain specificity in cognition and culture*. (p. 3-36). Cambridge University Press.

Horton, J.C. et Adams, D.L. (2005). The cortical column: a structure without a function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 837-862.

Hughes, C. et Cutting, A.L. (1999). Nature, nurture, and individual differences in early understanding of mind. *Psychological Science*, 10(5), 429-432.

Hutto, D.D. (2007). The narrative practice hypothesis: origins and applications of folk psychology. *Royal Institute of Philosophy Supplement*, 60, 43.

Jacobs, R.A. (1999). Computational studies of the development of functionally specialized neural modules. *Trends in Cognitive Sciences*, 3(1), 31-38.

Jenkins, A.C. et Mitchell, J.P. (2010). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, 20(2), 404-410.

Joseph, R.M. et Tager–Flushberg, H. (2004). The relationship of theory of mind and executive functions to symptom type and severity in children with autism. *Development and psychopathology*, 16(01), 137-155.

Jungé, J.A. et Dennett, D.C. (2010). Multi-use and constraints from original use. *Behavioral and Brain Sciences*, 33(04), 277-278.

Kaas, J.H. (2012). Evolution of columns, modules, and domains in the neocortex of primates. *Proceedings of the National Academy of Sciences*, 109(S1), 10655-10660.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience*, 3, 759-763.

Kanwisher, N. (2010). Functional specificity in the human brain : a window into the

functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25), 11163-11170.

Kanwisher, N., McDermott, J. et Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11), 4302-4311.

Kanwisher, N. et Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476), 2109-2128.

Kaplan, J.M. (2002). Historical evidence and human adaptations. *Philosophy of science*, 69(S3), S294-S304.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA : MIT press.

Klein, S.B., Cosmides, L., Tooby, J. et Chance, S. (2002). Decisions and the evolution of memory: multiple systems, multiple functions. Psychological review, 109(2), 306-329.

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: the Social Attribution Task. *Journal of Child psychology and Psychiatry*, 41(7), 831-846.

Kobayashi, C., Glover, G.H. et Temple, E. (2007). Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. *Neuropsychologia*, 45(7), 1522-1532.

Kobayashi Frank, C. et Temple, E. (2009). Cultural effects on the neural basis of theory of mind. *Progress in brain research*, 178, 213-223.

Koster-Hale, J. et Saxe, R. (2013). Theory of Mind : A Neural Prediction Problem. *Neuron*, 79(5), 836-848.

Krebs, J.R., Sherry, D.F., Healy, S.D., Perry, V.H. et Vaccarino, A.L. (1989). Hippocampal specialization of food-storing birds. *Proceedings of the National Academy of Sciences*, 86(4), 1388-1392.

Krill, A.L., Platek, S.M., Goetz, A.T. et Shackelford, T.K. (2007). Where Evolutionary Psychology meets Cognitive Neuroscience : A précis to Evolutionary Cognitive Neuroscience. *Evolutionary Psychology*, 5(1), 232-256.

Krubitzer, L. (2007). The magnificent compromise : cortical field evolution in mammals. *Neuron*, 56(2), 201-208.

Kurzban, R., Tooby, J. et Cosmides, L. (2001). Can race be erased? Coalitional

computation and social categorization. *Proceedings of the National Academy of Sciences*, 98(26), 15387-15392.

LeDoux, J.E. (2012). Evolution of human emotion: a view through fear. *Progress in brain research*, 195, 431.

Leise, E.M. (1990). Modular construction of nervous systems: a basic principle of design for invertebrates and vertebrates. *Brain Research Reviews*, 15(1), 1-23.

Leslie, A.M. (1992). Pretense, autism, and the theory-of-mind module. *Current Directions in Psychological Science*, 1(1), 18-21.

Leslie, A.M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. Dans L.A. Hirschfeld et S.A. Gelman (dir.), *Mapping the mind : Domain specificity in cognition and culture*. (p. 119-148), Cambridge, UK : Cambridge University Press.

Leslie, A.M., Friedman, O. et German, T.P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12), 528-533.

Leslie, A.M., German, T.P. et Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50(1), 45-85.

Leslie, A.M. et Thaiss, L. (1992). Domain specificity in conceptual development : Neuropsychological evidence from autism. *Cognition*, 43(3), 225-251.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australian Journal of Philosophy*, 50(3), 249-258.

Lewontin, R.C. (1979). Sociobiology as an adaptationist program. *Behavioral science*, 24(1), 5-14.

Lickliter, R. et Honeycutt, H. (2003). Developmental dynamics : toward a biologically plausible evolutionary psychology. *Psychological bulletin*, 129(6), 819-835.

Lieberman, D., Tooby, J. et Cosmides, L. (2007). The architecture of human kin detection. *Nature*, 445(7129), 727-731.

Liu, D., Wellman, H.M., Tardif, T. et Sabbagh, M.A. (2008). Theory of mind development in Chinese children : a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44(2), 523.

Livingstone, M.S. et Hubel, D.H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal of Neuroscience*, 7(11), 3416-3468.

Lloyd, E.A. (1999). Evolutionary psychology : The burdens of proof. *Biology and Philosophy*, 14(2), 211-233.

Lloyd, E.A. (2012). Units and Levels of Selection. *The Stanford Encyclopedia of Philosophy* (Winter 2012 éd.). http://plato.stanford.edu/archives/win2012/entries/selection-units/.

Lombardo, M.V., Baron-Cohen, S., Belmonte, M. et Chakrabarti, B. (2011). Neural endophenotypes for social behaviour in autism spectrum conditions. Dans J. Decety et J. T. Cacioppo (dir.), *Handbook of social neuroscience*. Oxford University Press.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T. et Baron-Cohen, S. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *Neuroimage*, 56(3), 1832-1838.

Love, A.C. (2007). Functional homology and homology of function : Biological concepts and philosophical consequences. *Biology & Philosophy*, 22(5), 691-708.

Ludwig, K. et Schneider, S. (2008). Fodor's challenge to the classical computational theory of mind. *Mind & Language*, 23(1), 123-143.

Machery, E. (2007). Massive modularity and brain evolution. *Philosophy of Science*, 74(5), 825-838.

Machery, E. (2011). Developmental disorders and cognitive architecture. Dans A. De Block et P. Adriaens (dir.), *Maladapting minds* (p. 91-116). Oxford : Oxford University Press.

Machery, E. (2014). In Defense of Reverse Inference. *The British Journal for the Philosophy of Science*, 65(2), 251-267.

Machery, E. (à venir). Discovery and confirmation in evolutionary psychology. Dans J. Prinz (dir.), *Oxford Handbook of Philosophy of Psychology*. Oxford University Press.

Machery, E. et Barrett, H.C. (2006). Essay Review: Debunking Adapting Minds. *Philosophy of Science*, 73(2), 232-246.

MacLean, E.L. et Hare, B. (2012). Bonobos and chimpanzees infer the target of another's attention. *Animal Behaviour*, 83(2), 345-353.

MacLean, E.L., Matthews, L.J., Hare, B.A., Nunn, C.L., Anderson, R.C., Aureli, F., Brannon, E.M., Call, J., Drea, C.M. et Emery, N.J. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal cognition*, 15(2), 223-

238.

Mahon, B.Z. et Cantlon, J.F. (2011). The specialization of function : Cognitive and neural perspectives. *Cognitive neuropsychology*, 28(3-4), 147-155.

Mameli, M. (2001). Modules and mindreaders. *Biology and Philosophy*, 16(3), 377-393.

Mantini, D., Corbetta, M., Romani, G.L., Orban, G.A. et Vanduffel, W. (2013). Evolutionarily Novel Functional Networks in the Human Brain? *The Journal of Neuroscience*, 33(8), 3259-3275.

Marcus, G.F. (2003). *The algebraic mind : Integrating connectionism and cognitive science*. Cambridge, MA : MIT Press.

Marcus, G.F. (2006). Cognitive architecture and descent with modification. *Cognition*, 101(2), 443-465.

Marcus, G.F. (2009). How *does* the mind work? Insights from biology. *Topics in Cognitive Science*, 1(1), 145-172.

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*. New York : W.H. Freeman and Company.

Mars, R.B., Sallet, J., Neubert, F.-X. et Rushworth, M.F. (2013). Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex. *Proceedings of the National Academy of Sciences*, 110(26), 10806-10811.

Mars, R.B., Sallet, J., Schüffelgen, U., Jbabdi, S., Toni, I. et Rushworth, M.F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral cortex*, 22(8), 1894-1903.

Marshall, J.C. (1984). Multiple perspectives on modularity. *Cognition*, 17(3), 209-242.

Martin, A. et Santos, L.R. (2014). The origins of belief representation : Monkeys fail to automatically represent others' beliefs. *Cognition*, 130(3), 300-308.

Marzke, M.W. et Marzke, R.F. (2000). Evolution of the human hand : approaches to acquiring, analysing and interpreting the anatomical evidence. *Journal of*

anatomy, 197(01), 121-140.

Matthen, M. (2007). Defining vision : what homology thinking contributes. *Biology & Philosophy*, 22(5), 675-689.

Mayr, E. (1960). The emergence of evolutionary novelties. Dans S. Tax (dir.), *Evolution After Darwin*, Vol. II (p. 349-380). Chicago : University Chicago Press.

McClamrock, R. (1993). Functional analysis and Etiology. *Erkenntnis*, 38(2), 249-260.

Meunier, D., Lambiotte, R. et Bullmore, E.T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 200(4), 1-11.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, 18(2), 262-271.

Mitchell, J.P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1309-1316.

Morgan, M.H. et Carrier, D.R. (2013). Protective buttressing of the human fist and the evolution of hominin hands. *The Journal of experimental biology*, 216(2), 236-244.

Mountcastle, V. B., (1978). An organizing principle for cerebral function : The unit module and the distributed system, Dans M. Edelman et V. Mountcastle (dir.), *The mindful brain* (p. 7-50). Cambridge, MA: MIT Press.

Mundale, J. (2003). Concepts of localization : Balkanization in the brain. *Brain and Mind*, 3(3), 313-330.

Neuberg, S.L., Kenrick, D.T., et Shaller, M. (2010). Evolutionary social cognition. Dans S.T. Fiske, D. Gilbert et G. Lindzey(dir.), *Handbook of Social cognition*. (p. 761-796). New York : John Wiley & Sons.

Nichols, S. (à venir). Mindreading and the Philosophy of Mind. Dans J. Prinz (dir.), *The Oxford Handbook on Philosophy of Psychology*. New York : Oxford University Press.

Nichols, S. et Stich, S.P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* New York : Oxford University Press.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. et Clark, A.G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8(11),

857-868.

Öhman, A. et Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, 108(3), 483.

Olshausen, B.A. et Field, D.J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4), 481-487.

Onishi, K.H. et Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.

Parnas, D.L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12), 1053-1058.

Penn, D.C., Holyoak, K.J. et Povinelli, D.J. (2008). Darwin's mistake : Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2), 109-129.

Perlman, S.B., Vander Wyk, B.C. et Pelphrey, K.A. (2010). Brain mechanisms in the typical and atypical development of social cognition. *Developmental social cognitive neuroscience*, 99-124.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W. et Ladurner, G. (2006). Thinking of mental and other representations : The roles of left and right temporo-parietal junction. *Social neuroscience*, 1(3-4), 245-258.

Perner, J. et Leekam, S. (2008). The curious incident of the photo that was accused of being false : Issues of domain specificity in development, autism, and brain imaging. *The Quarterly Journal of Experimental Psychology*, 61(1), 76-89.

Pinker, S. (1994). *The Language Instinct*. New York : William Morrow and Company.

Pinker, S. (1997). *How the mind works*. New York : Norton.

Pinker, S. (2005a). A reply to Jerry Fodor on how the mind works. *Mind & Language*, 20, 33-38.

Pinker, S. (2005b). So how does the mind work? *Mind & Language*, 20(1), 1-24.

Pinker, S. et Bloom, P. (1990). Natural selection and natural language. *Behavioral and Brain Sciences*, 13(4), 707-784.

Pinker, S. et Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95(2), 201-236.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data?

Trends in cognitive sciences, 10(2), 59-63.

Premack, D. et Premack, A.J. (1995). Intention as psychological cause. Dans D. Sperber, D. Premack et A.J. Premack (dir.), *Causal cognition : A multidisciplinary debate* (p. 185-199). Oxford University Press.

Premack, D. et Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(04), 515-526.

Preuss, T.M. (2001). The discovery of cerebral diversity : an unwelcome scientific revolution. Dans D. Falk et K. Gibson (dir.), *Evolutionary anatomy of the primate cerebral cortex* (p. 138-164). Cambridge : Cambridge University Press.

Preuss, T.M. (2011). The human brain : rewired and running hot. *Annals of the New York Academy of Science*s, 1225(S1), E182-E191.

Preuss, T.M. (2012). Human brain evolution : From gene discovery to phenotype discovery. *Proceedings of the National Academy of Sciences*, 109(S1), 10709-10716.

Prinz, J. (2006). Is the mind really modular? Dans R.J. Stainton (dir.), *Contemporary debates in cognitive science* (p. 22-36). Blackwell.

Quartz, S.R. et Sejnowski, T.J. (1997). The neural basis of cognitive development : A constructivist manifesto. *Behavioral and brain sciences*, 20(4), 537-556.

Ramus, F. (2006). Genes, brain, and cognition : A roadmap for the cognitive scientist. *Cognition*, 101(2), 247-269.

Ritchie, J.B. et Carruthers, P. (2010). Massive modularity is consistent with most forms of neural reuse. *Behavioral and Brain Sciences*, 33(04), 289-290.

Robbins, P. (2010). Modularity of mind. *The Stanford Encyclopedia of Philosophy*. (Summer 2010 éd.) http://plato.stanford.edu/archives/sum2010/entries/modularity-mind/.

Ross, L.A. et Olson, I.R. (2010). Social cognition and the anterior temporal lobes. *Neuroimage*, 49(4), 3452-3462.

Rushworth, M.F., Mars, R.B. et Sallet, J. (2013). Are there specialized circuits for social cognition and are they unique to humans? *Current opinion in neurobiology*, 23(3), 436-442.

Samson, D. et Apperly, I.A. (2010). There is more to mind reading than having theory of mind concepts: New directions in theory of mind research. *Infant and Child Development*, 19(5), 443-454.

Samson, D., Apperly, I.A., Chiavarino, C. et Humphreys, G.W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature neuroscience*, 7(5), 499-500.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *The British Journal for the Philosophy of Science*, 49(4), 575-602.

Samuels, R. (2000). Massively modular minds : Evolutionary psychology and cognitive architecture. Dans P. Carruthers (dir.), *Evolution and the human mind: Modularity, language and meta-cognition* (p. 13-46). Cambridge University Press.

Samuels, R. (2005). The complexity of cognition : tractability arguments for massive modularity". Dans P. Carruthers, S. Laurence et S. Stich (dir.), *The Innate Mind: Structure and Contents* (p. 107-121). Oxford University Press..

Samuels, R. (2006). Is the mind massively modular? Dans R.J. Stainton (dir.), *Contemporary debates in cognitive science* (p. 37-56). Blackwell.

Samuels, R. (2010). Classical computationalism and the many problems of cognitive relevance. *Studies in History and Philosophy of Science Part A*, 41(3), 280-293.

Santos, L.R., Flombaum, J.I. et Phillips, W. (2007). The Evolution of Human Mindreading: How Nonhuman Primates Can Inform Social Cognitive Neuroscience. Dans S. Platek, J. Keenan et T. Shackelford (dir.), *Evolutionary cognitive neuroscience* (p. 433-456). Cambridge, MA : MIT Press.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in cognitive sciences*, 9(4), 174-179.

Saxe, R. (2006). Uniquely human social cognition. *Current opinion in neurobiology*, 16(2), 235-239.

Saxe, R. (2009). The neural evidence for simulation is weaker than I think you think it is. *Philosophical studies*, 144(3), 447-456.

Saxe, R. (2010). The right temporo-parietal junction: A specific brain region for thinking about thoughts. Dans A. Leslie et T. German (dir.), *Handbook of theory of mind* (p. 1-35).

Saxe, R., Carey, S. et Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87-124.

Saxe, R. et Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4), 1835-

1842.

Saxe, R. et Powell, L.J. (2006). It's the Thought That Counts Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, 17(8), 692-699.

Saxe, R., Schulz, L.E. et Jiang, Y.V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience*, 1(3-4), 284-298.

Saxe, R. et Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43(10), 1391-1399.

Schaller, M., Park, J.H. et Kenrick, D.T. (2007). Human evolution and social cognition. Dans R.I.M. Dunbar et L. Barrett (dir.), *Oxford handbook of evolutionary psychology* (p. 491-504). Oxford University Press.

Schenker, N.M., Desgouttes, A.-M. et Semendeferi, K. (2005). Neural connectivity and cortical substrates of cognition in hominoids. *Journal of Human Evolution*, 49(5), 547-569.

Schlosser, G. et Wagner, G.P. (2004). *Modularity in development and evolution*. University of Chicago Press.

Schmitt, D.P. et Pilcher, J.J. (2004). Evaluating evidence of psychological adaptation : How do we know one when we see one? *Psychological Science*, 15(10), 643-649.

Scholl, B.J. (1997). Neural constraints on cognitive modularity? *Behavioral and Brain Sciences*, 20(04), 575-576.

Scholl, B.J. et Leslie, A.M. (1999). Modularity, development and 'theory of mind'. *Mind & Language*, 14(1), 131-153.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E.N. et Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*, 4(3), e4869.

Schulz, A.W. (2011). Simulation, simplicity, and selection: an evolutionary perspective on high-level mindreading. *Philosophical studies*, 152(2), 271-285.

Schulz, A.W. (2012) Heuristic Evolutionary Psychology. Dans K. Plaisance et T. Reydon (dir.), *Philosophy of behavioral biology* (p. 217-234). Berlin : Springer.

Segal, G. (1996). The modularity of theory of mind. Dans P. Carruthers et P. K. Smith (dir.), *Theories of Theories of Mind* (p. 141-157). Cambridge, UK : Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. Dans H. Feigl et M. Scriven (dir.), *Minnesota Studies in the Philosophy of Science* (Vol. 1, p. 253-329). Minneapolis, MN : University of Minnesota Press.

Seok, B. (2006). Diversity and unity of modularity. *Cognitive science*, 30(2), 347-380.

Shallice, T. (1984). More functionally isolable subsystems but fewer "modules"? *Cognition*, 17(3), 243-252.

Shallice, T. (1988). *From neuropsychology to mental structure*. New York : Cambridge University Press.

Shapiro, L. et Epstein, W. (1998). Evolutionary theory meets cognitive psychology: A more selective perspective. *Mind & Language*, 13(2), 171-194.

Shapiro, L.A. (2001). Mind the Adaptation. Dans D. Walsh (dir.), *Evolution, Naturalism and Mind.* (p. 23-41). Cambridge, UK : Cambridge University Press.

Sherry, D.F. et Schacter, D.L. (1987). The evolution of multiple memory systems. *Psychological review*, 94(4), 439.

Shettleworth, S.J. (2000). Modularity and the evolution of cognition. Dans C. Heyes et L. Huber (dir.), *The Evolution of Cognition* (p. 43-69). Cambridge, MA : MIT Press.

Shettleworth, S.J. (2009). *Cognition, evolution, and behavior*. Oxford University Press.

Simon, H.A. (1962). The architecture of complexity. *General systems*, 10(1965), 63-76.

Simpson, J.A. et Campbell, L. (2005). Methods of Evolutionary Sciences. Dans D. Buss (dir.), *The Handbook of Evolutionary Psychology* (p. 119-144). New Jersey : Wiley.

Spelke, E. et Dehaene, S. (1999). Biological foundations of numerical thinking: Response to TJ Simon (1999). *Trends in Cognitive Sciences*, 3(10), 365-366.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. Dans L.A. Hirschfeld et S.A. Gelman (dir.), *Mapping the mind: Domain Specificity in Cognition and Culture*. (p. 39-67). New York : Cambridge University Press.

Sperber, D. (2002). In defense of massive modularity. Dans E. Dupoux (dir.), *Language, brain and cognitive development: Essays in honor of Jacques Mehler*

(p. 47–57). Cambridge, MA: MIT Press.

Sperber, D. (2005). Modularity and relevance : How can a massively modular mind be flexible and context-sensitive? Dans P. Carruthers, S. Laurence et S. Stich (dir.), *The innate mind : Structure and content* (p. 53-68). Oxford, England : Oxford University Press.

Sperber, D., Cara, F. et Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57(1), 31-95.

Sperber, D. et Girotto, V. (2003). Does the selection task detect cheater-detection? Dans K. Sterelny et J. Fitness (dir.), *From Mating to Mentality : Evaluating Evolutionary Psychology* (p. 197-226). Macquarie University Series in Cognitive Psychology.

Sripada, C.S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, 48(1), 232-238.

Sterelny, K. et Griffiths, P.E. (1999). *Sex and death : An introduction to philosophy of biology*. Chicago/London : University of Chicago Press.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta psychologica*, 106(1), 147-246.

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

Stich, S. et Nichols, S. (2003). Folk psychology. Dans S. Stich et T. Warfield (dir.), *The Blackwell guide to philosophy of mind* (p. 235-255). Oxford : Blackwell.

Stone, V.E. (2005). Theory of mind and the evolution of social intelligence. Dans J. Cacciopo (dir.), *Social neuroscience: People thinking about people* (p. 103-130). Boston : MIT Press.

Stone, V.E. (2007). An Evolutionary perspective on domain specificity in social intelligence. Dans E. Harmon-Jones et P. Winkielman (dir.), *Social Neuroscience : Integrating Biological and Psychological Explanations of Social Behavior* (p. 316-349). New York/London : Guilford Press.

Stone, V.E., Baron-Cohen, S. et Knight, R.T. (1998). Frontal lobe contributions to theory of mind. *Journal of cognitive neuroscience*, 10(5), 640-656.

Stone, V.E., Cosmides, L., Tooby, J., Kroll, N. et Knight, R.T. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences*, 99(17), 11531-11536.

Stone, V.E. et Gerrans, P. (2006). What's domain-specific about theory of mind? *Social Neuroscience*, 1(3-4), 309-319.

Stotz, K.C. et Griffiths, P.E. (2002). Dancing in the dark : Evolutionary psychology and the problem of design. Dans F. Rauscher et S. Scher (dir.), *Evolutionary Psychology : Alternative Approaches* (p. 135-160). Dortrecht: Kluver.

Striedter, G.F. et Northcutt, R.G. (1991). Biological hierarchies and the concept of homology. *Brain, Behavior and Evolution*, 38(4-5), 177-189.

Suddendorf, T. (2008). Explaining human cognitive autapomorphies. *Behavioral and Brain Sciences*, 31(02), 147-148.

Thornhill, R. (2007). Comprehensive knowledge of human evolutionary history requires both adaptationism and phylogenetics. Dans S. W. Gangestad et J. A. Simpson (dir.), *The evolution of mind: Fundamental questions and controversies* (p. 31-37). New York: Guilford Press.

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

Tomasello, M., Melis, A.P., Tennie, C., Wyman, E. et Herrmann, E. (2012). Two key steps in the evolution of human cooperation. *Current Anthropology*, 53(6), 673-692.

Tooby, J. et Cosmides, L. (1989). Adaption Versus Phylogeny: The Role of Animal Psychology in The Study of Human Behavior. *International Journal of Comparative Psychology*, 2(3), 175-188.

Tooby, J. et Cosmides, L. (1992). The psychological foundations of culture. Dans J. H. Barkow, L. Cosmides et J. Tooby (dir.), *The Adapted Mind : Evolutionary Psychology and the Generation of Culture* (p. 19-136). NY/Oxford : Oxford University Press.

Tooby, J. et Cosmides, L. (1995a). The language of the eyes as an evolved language of mind. Dans S. Baron-Cohen (dir.), *Mindblindness : An Essay on Autism and Theory of Mind* (p. XI-XVIII). Cambridge, MA : MIT Press.

Tooby, J. et Cosmides, L. (1995b). Mapping the Evolved Functional Organization of Mind and Brain. Dans M. Gazzaniga (dir.), *The cognitive neurosciences*. Cambridge, MA : MIT Press.

Tooby, J. et Cosmides, L. (2005). Conceptual foundations of evolutionary psychology. Dans D. Buss (dir.), *The handbook of evolutionary psychology* (p. 5-67). New Jersey : Wiley.

Tooby, J., Cosmides, L. et Barrett, H.C. (2005). Resolving the debate on innate ideas: Learnability constraints and the evolved interpenetration of motivational and conceptual functions. Dans P. Carruthers, S. Laurence et S. Stich (dir.), *The innate mind : Structure and content* (p. 305-337). New York : Oxford University Press.

Uttal, W.R. (2001). *The new phrenology : the limits of localizing cognitive processes in the brain*. Cambridge, MA : MIT Press.

Uttich, K. et Lombrozo, T. (2010). Norms inform mental state ascriptions : A rational explanation for the side-effect effect. *Cognition*, 116(1), 87-100.

Van Orden, G.C. et Kloos, H. (2003). The module mistake. *Cortex*, 39(1), 164-166.

Van Orden, G.C., Pennington, B.F. et Stone, G.O. (2001). What do double dissociations prove? *Cognitive Science*, 25(1), 111-172.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human brain mapping*, 30(3), 829-858.

Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *Neuroimage*, 54(2), 1589-1599.

Van Overwalle, F. et Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, 48(3), 564-584.

Vonk, J. et Povinelli, D.J. (2006). Similarity and difference in the conceptual systems of primates : the Unobservability hypothesis. Dans T. Zentall, E.A. Wasserman (dir.), *Comparative Cognition : Experimental Explorations of Animal Intelligence* (p. 363-387). Oxford, UK : Oxford University Press.

Wagner, W. et Wagner, G.P. (2003). Examining the modularity concept in evolutionary psychology: the level of genes, mind, and culture. *Journal of Cultural and Evolutionary Psychology*, 1(3), 135-165.

Wang, Y.W., Lin, C.D., Yuan, B., Huang, L., Zhang, W.X. et Shen, D.L. (2010). Person perception precedes theory of mind: an event related potential analysis. *Neuroscience*, 170(1), 238-246.

Wellman, H.M., Cross, D. et Watson, J. (2001). Meta-analysis of theory-of-mind development : the truth about false belief. *Child development*, 72(3), 655-684.

Williams, G.C. (1966). *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton University Press.

Wimmer, H. et Perner, J. (1983). Beliefs about beliefs : Representation and

constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128.

Woodward, J. et Cowie, F. (2004). The mind is not (just) a system of modules shaped (just) by natural selection. Dans Hitchcock, C. (dir.), *Contemporary debates in the philosophy of science* (p. 312-334). Wiley-Blackwell.

Woolfe, T., Want, S.C., et Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development*, 73, 768–778.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 34(4), 633-668.

Xia, H., Wu, N. et Su, Y. (2012). Investigating the genetic basis of theory of mind (ToM): the role of catechol-O-methyltransferase (COMT) gene polymorphisms. *PloS ONE*, 7(11), e49768.

Zaitchik, D. (1990). When representations conflict with reality: the preschooler's problem with false beliefs and "false" photographs. *Cognition*, 35(1), 41-68.

Zaki, J., Hennigan, K., Weber, J. et Ochsner, K.N. (2010). Social cognitive conflict resolution: contributions of domain-general and domain-specific neural systems. *The Journal of Neuroscience*, 30(25), 8481-8488.

Zaki, J. et Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences*, 1167, 16-30.

Zaki, J. et Ochsner, K. (2011). Reintegrating the Study of Accuracy Into Social Cognition Research, *Psychological Inquiry*, 22(3), 159-182.

Zaki, J., Weber, J., Bolger, N. et Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences*, 106(27), 11382-11387.

Zawidski, T. et Bechtel, W.P. (2004). Gall's legacy revisited: Decomposition and localization in cognitive neuroscience. Dans C.E. Erneling et D.M. Johnson (dir.), *Mind As a Scientific Object.* Oxford University Press.

Zawidzki, T.W. (2008). The function of folk psychology : Mind reading or mindshaping? *Philosophical explorations*, 11(3): 193-209.

Zürcher N.R., Donnelly N., Rogier O., Russo B., Hippolyte L., et al. (2013) It's All in the Eyes : Subcortical and Cortical Activation during Grotesqueness Perception in Autism. *PLoS ONE*, 8(1), e54313.