## Approches psychologiques, cognitives et neurobiologiques de l'utilisation d'outils et de l'apprentissage sensorimoteur

Dans cette partie, nous avons sélectionné quelques éléments de la littérature de ces différents domaines proposant soit des résultats fondamentaux, soit des théories portant sur des mécanismes qui nous ont semblé pertinents pour les travaux de cette thèse. Il ne s'agira bien souvent que d'un bref aperçu ne prétendant nullement à l'exhaustivité, mais permettant de donner des éléments clés pour la modélisation.

Afin de délimiter au mieux le champs de notre recherche, nous commencerons par nous intéresser spécifiquement à la question de l'utilisation d'outils (Sec. 1.1.1) en tentant de donner un bref aperçu des problématiques dites bas niveaux et hauts niveaux auxquelles la littérature rattache cette question. Ce bref panorama nous permettra de parcourir ces problématiques dans les sections qui suivent.

Puisque la question de l'encodage sensorimoteur est première et constitue l'axe choisi pour aborder cette large question, nous extrairons ensuite un ensemble de points qui nous ont paru clés dans la théorie des contingences sensorimotrices (Sec. 1.1.2). Dans la continuité de cette question fondamentale du lien entre senseur et moteur, nous donnerons également un bref aperçu du principe idéomoteur (Sec. 1.1.3).

Afin d'embrasser plus largement la problématique de l'utilisation d'outil, suite à ces considérations fondamentales nous observerons rapidement les questions soulevées par la capacité à effectuer des séquences d'actions (Sec. 1.1.4), l'une des caractéristiques de l'utilisation d'outil.

Enfin nous examinerons diverses théories psychologiques basées sur des observations neurobiologiques, non spécifiques à l'utilisation d'outil bien qu'on les y rattache parfois, dont la transversalité à de nombreuses thématiques est riche en enseignements concernant certains mécanismes clés du cerveau. Nous traiterons ainsi des affordances, des neurones miroirs, et de la théorie de la simulation (Sec. 1.1.5).

### 1.1.1 Utilisation d'outils

La question de l'utilisation d'outils est à la croisée de différentes problématiques. C'est une question qui met en jeu une multitude de domaines, et les problématiques dans lesquelles elle nous plonge reflètent la nécessité d'une approche pluridisciplinaire pour apporter une réponse à la mesure du problème. La question de l'aptitude à utiliser des outils fait actuellement l'objet de recherches non seulement en robotique, pour permettre aux robots de satisfaire de nouvelles tâches, mais aussi en psychologie développementale pour comprendre les étapes nécessaires à la maîtrise de cette faculté et les mécanismes sous-jacents qui sont impliqués. De nombreuses études portent également sur les aires du cerveau mises en jeu dans l'obtention de cette faculté, et enfin, surplombant l'ensemble de ces aspects, la question philosophique du rapport entre l'utilisation d'outils et ce qui définit l'homme en tant qu'homme, question qui passe notamment par la question du langage, vient donner toute sa profondeur et sa portée à la problématique qui nous occupe.

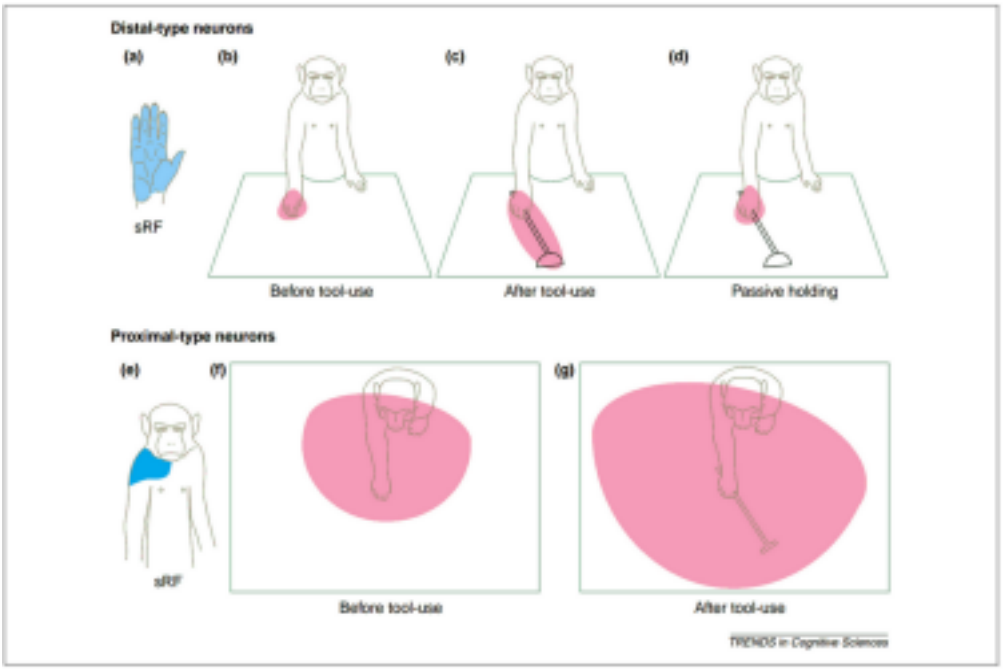


**FIGURE 1.1 –** Figure illustrant l'effet de l'utilisation d'outils sur le schéma corporel (position de la main, espace atteignable). Issu de [Maravita and Iriki, 2004]

Si, bien entendu, nous n'avons pas prétention à couvrir l'ensemble de ces domaines, et encore moins à y proposer une solution viable, c'est en gardant l'ensemble de ces questions à l'esprit que la réflexion a tenté d'être portée, afin d'en respecter l'étendue.

Dans cette section nous nous en tiendrons cependant à un résumé de la problématique d'un point de vue essentiellement psychologique.

Dans ce que l'on pourrait classer parmi les problématiques dites bas niveaux, l'utilisation d'outils suppose une extension du schéma corporel : des résultats montrent en effet, chez les macaques, que des neurones bimodaux répondant à la vision et à la proprioception réagissent différemment après l'utilisation active d'outils. En effet, des neurones codant la position de la main ont une activité étendue à la position de l'outil, et l'espace atteignable se trouve lui aussi étendu (voir figure 1.1). Ces résultats semblent indiquer qu'à très bas niveau l'outil est susceptible d'être perçu comme une extension du corps, et encode donc une extension du schéma corporel, différente selon les outils utilisés (voir les travaux de [Maravita and Iriki, 2004; Arbib et al., 2009] et de [Hoffmann et al., 2010] pour une analyse critique de la question et de la manière dont un tel schéma corporel peut être encodé dans un robot). Notons que la question reste ouverte de savoir dans quelle mesure un modèle interne de l'outil est appris, et dans quelle mesure c'est le modèle du bras lui-même qui est adapté [Kluzik et al., 2008].

L'aspect "bas niveau" est ici à comprendre comme étant fondamental, au sens où la question d'une telle plasticité dans l'apprentissage sera le fondement de toutes constructions ultérieures, que l'on pourrait alors qualifier de plus "haut niveau", quand en définitif celles-ci dépendront principalement des caractéristiques de l'encodage sensorimoteur choisi. Or la question de ce qui doit être appris rejoint la question plus générale du lien entre senseurs et moteurs : il s'agit

d'une question non tranchée, et flirtant avec la philosophie en ce que prendre position dessus définit globalement une approche, parmi d'autres, au sein de différents courants de pensées faisant toujours l'objet de débats (voir par exemple [Jacob and Jeannerod, 2005; O'Regan and Block, 2012]). La question du lien entre senseurs et moteurs sera abordée dans les deux sections suivantes, en ayant donc le partie pris de ne nous intéresser dans cet état de l'art qu'à la théorie des contingences sensorimotrices, et au principe idéomoteur.

Parmi les caractéristiques qualifiées de "hauts niveaux", l'utilisation d'outils est aussi liée à des capacités cognitives de plus hauts niveaux (voir [Guerin et al., 2013]), et à des capacités d'abstraction telles que celles permettant de faire des séquences ([McCarty et al., 1999; Johnson, 2000; Johnson-Frey, 2004]), et est fortement liée aux problèmes du langage ([Stout and Chaminade, 2009; Cangelosi et al., 2010; Arbib, 2011; Steele et al., 2012]).

Dans les expériences menées par Fagard *et al.*, des enfants doivent utiliser un outil en forme de râteau afin d'atteindre des jouets hors de portée. Les résultats indiquent que ce n'est seulement qu'entre le $16^{\text{ème}}$ et le $20^{\text{ème}}$ mois que les enfants commencent à intentionnellement essayer de rapprocher l'objet avec l'outil, ce qui suggère qu'une véritable compréhension de l'utilisation d'outils n'est pas acquise avant cet âge [Rat-Fischer et al., 2012]. Notons de plus que les performances de l'enfant se trouve améliorées lorsque celui-ci comprend les intentions du démonstrateur lors d'un apprentissage par observation de cette tâche [Esseily et al., 2013].

Parmi les capacités notables que l'enfant développe avant ses 18 mois, on peut noter le comportement de moyen en vue d'une fin ("means-end") qui apparaît autour du $8^{\text{ème}}$ au $12^{\text{ème}}$ mois. Celui-ci implique l'exécution planifiée d'une séquence d'étapes permettant d'accomplir un but, et apparaît dans des situations dans lesquelles un obstacle doit être levé pour satisfaire le but [Willatts, 1999]. Ce comportement constitue une étape très importante et nécessaire pour l'utilisation d'outil, mais il est intéressant de noter que la capacité à planifier, et à faire des séquences, n'est pas suffisante pour l'utilisation d'outil que l'on observe plus tard, malgré l'aspect haut niveau des capacités qui sont pourtant requises.

Ainsi, même si la planification et l'exécution d'une séquence d'étapes permettant d'atteindre un but est un comportement complexe et fortement lié à l'utilisation d'outil, de tels comportements (requis pour ceux de "means-end") se développent avant l'apparition de celui d'utilisation d'outil, et ne sont donc pas suffisants. Notons également que l'apraxie idéatoire montre qu'une décorrélation est possible entre la capacité à effectuer certains mouvements, et la capacité à organiser ces mouvements en séquences visant un but précis, et ceci impacte la capacité à utiliser des outils [Osiurak et al., 2010]. La question plus spécifique de la capacité à faire des séquences sera ainsi abordée dans la section 1.1.4.

D'autres types de travaux ([Stout and Chaminade, 2007]), basés sur l'imagerie cérébrale, ont eux montré que, sur des sujets apprenant à faire des outils en pierre, la conceptualisation abstraite ou la planification stratégique n'est pas ce qui est le plus important quand il s'agit de l'utilisation d'outil, mais semble indiquer qu'il s'agit plutôt d'adaptation sensorimotrice, et de perception d'affordances. Cette idée est également développée dans [Osiurak et al., 2010], et nous explorerons des théories relatives aux affordances en section 1.1.5.

Pour finir nous noterons que, dans le cadre de cette thèse, la problématique de l'utilisation d'outils nous semble une bonne illustration de la problématique de l'"open-ended" telle qu'on la retrouve en robotique (et que nous développerons en section 1.2.6). En effet, un des cas qui semble typiquement refléter une capacité à utiliser un outil est lorsque celui-ci est entièrement nouveau (quelles que soient les différentes modalités qu'il sollicite) pour le robot : car autrement

il peut suffire, pour un roboticien, de l'encoder "en dur". Or si ça peut être le cas pour le bras ou la main, qui peuvent être interprétés comme les premiers outils dont nous disposons, nous pensons plutôt que c'est *d'une même manière* que l'usage d'un quelconque outil fait l'objet d'un apprentissage. Ainsi un robot aura à découvrir sans cesse de nouvelles aptitudes impactant ce qu'il est à même de faire, et donc, potentiellement, ayant également un impact sur ses précédents apprentissages sensorimoteurs. Il devra également faire appel à cette nouvelle aptitude en fonction du contexte, pour résoudre des tâches connues ou inconnues à l'avance. Notons que la capacité à mobiliser des connaissances liées à l'utilisation d'outil, au cours d'une tâche où l'outil n'est pas explicitement requis, sera notamment l'objet de cette thèse.

### 1.1.2 Théorie des contingences sensorimotrices

L'approche sensorimotrice se caractérise par un questionnement radical sur ce qu'est une perception. Elle cherche à caractériser précisément ce qui est en jeu lorsque, par exemple, nous percevons la couleur rouge, ou encore lorsque nous pressons une éponge. Elle s'intéresse donc à la question des qualia, c'est-à-dire la manière dont les choses nous apparaissent, les propriétés subjectives de l'expérience accessibles à la conscience mais non communicable autrement que par analogie avec autrui (ceux-ci diffèrent donc des représentations, ou des caractéristiques fonctionnelles).

Pour J.K. O'Regan et A. Noé [O'Regan and Noë, 2001], les explications purement neuronales (comme la conductance, la fréquence d'excitation ou même le réseau auquel appartient un neurone) ne peuvent rendre compte de l'expérience de ces qualia, au sens où cela donne lieu à un saut explicatif ("explanatory gap") entre ce qu'on appelle les corrélats neuronaux et l'expérience hautement subjective des qualia. Ils rejettent donc l'idée que l'explication des qualia réside dans ces corrélats neuronaux, ou encore que ceux-ci puissent tenir lieu de représentations internes agissant comme des symboles représentant nos perceptions.

Selon ces auteurs, l'expérience des qualia ne peut être réduite à un état, car elle provient d'une activité, ou encore d'une manière d'agir. Le rôle du cerveau n'est pas de générer une perception, mais de permettre l'*interaction* à laquelle nous donnerons le nom de perception.

Ainsi percevoir n'est pas un état, mais une activité. Celle-ci est basée sur la connaissance des contingences sensorimotrices, à savoir les liens observés entre les sensations et les activités motrices. La perception de la mollesse sera par exemple basée sur l'activité d'écraser une éponge (voir fig. 1.2). Ces liens sont appris par l'expérience : le cerveau analyse les conséquences des mouvements qu'il commande sur les sensations qu'il reçoit en retour. La boucle complète entre commandes motrices (générées par le cerveau), monde physique, et sensations (recueillies en dernier lieu par le cerveau) est requise dans cet apprentissage. Une fois les lois des contingences sensorimotrices maîtrisées, un individu est à même de percevoir car il est à même de s'appuyer sur cette connaissance pour explorer le monde : il ne s'agit plus alors d'une succession de processus passifs, mais de la capacité à modifier activement ses sensations. Pour le dire autrement, la passivité intrinsèque des sensations (de part leur nature de capteurs et non d'actionneurs) est compensée par la connaissance des lois régissant leurs variations, lesquelles sont alors vécues comme perceptions.

Afin de mieux comprendre ce qui est en jeu, nous nous proposons de différencier ici quatre niveaux d'abstractions (allant du plus bas au plus haut niveau) en détaillant ce qui nous semble

être les éléments clés de la théorie des contingences sensorimotrices, dans le cadre de cette thèse : newpage

1 Le premier niveau est celui des informations sensorielles réelles (sur lesquelles on ne peut tricher), à savoir l'impact de la réalité physique sur les senseurs (comme la stimulation rétinienne par exemple). Nous pouvons supposer que ce niveau est entièrement passif : il s'agit de sensations passives, à l'opposé de la perception qui sera, elle, active. (Notons qu'ici nous qualifierons de "réel" ce qui se rapporte à la réalité physique objective.).

2 Le second niveau est celui des données qui seront traitées pas les lois. Notons que ce niveau n'est pas explicitement décrit en tant que tel par les auteurs, mais nous supposons son existence de leur explication aux phénomènes des rêves, ou des hallucinations. Dans [O'Regan and Block, 2012], O'Regan propose deux explications à ces phénomènes : l'une des deux (notons que l'autre sera expliquée au niveau 4) est que le cerveau se retrouve parfois comme piégé, et ce pour différentes raisons, dans un état similaire à celui dans lequel il peut être lorsqu'une interaction réelle a lieu. Ceci a pour effet d'entraîner l'activation de lois non reliées au contexte réel courant. De fait, cet "état", en activant l'une ou l'autre loi sensorimotrice, peut ne pas découler du seul premier niveau. C'est pourquoi nous proposons l'existence, d'un point de vue fonctionnel (peu importe ici sa nature réelle), d'un niveau intermédiaire, susceptible d'être alors le lieu de la "tricherie" opérée sur le cerveau durant une hallucination ou un rêve.

3 Le troisième niveau, le plus important, est celui de l'apprentissage des lois régissant les contingences sensorimotrices. Le phénomène de perception est caractérisé par la loi sensorimotrice qui s'applique (que celle-ci soit réelle ou pas, comme dans le cas du rêve). Les quatre points qui suivent décrivent plus spécifiquement les propriétés de ce niveau.

**FIGURE 1.2 –** Notre perception de la mollesse ne provient pas de l'excitation d'un ensemble de neurones, mais des propriétés régissant les lois sensorimotrices qui s'appliquent lors de l'interaction entre la main qui presse et l'éponge (sensation des doigts qui s'enfoncent, *etc.*). Schéma issu de [O'Regan, 2011].

**Apprentissage de la structure du changement :** Une contingence sensorimotrice est une dépendance, observée durant l'interaction avec le monde, entre des senseurs et des moteurs. L'apprentissage des lois régissant ces contingences sensorimotrices consiste en l'apprentissage de la structure du changement qui est observé lors de l'activation motrice. En appelant $S$ l'ensemble des toutes les entrées sensorielles, $M$ l'ensemble de toutes les commandes motrices, et $E$ l'ensemble de tous les états possibles de l'environnement, Philipona *et al.* [Philipona et al., 2003] en déduisent la relation $S = \psi(M, E)$ entre moteurs $M$, senseurs $S$ et états $E$ du monde. Les auteurs proposent de ne pas avoir un apprentissage du type $M = \phi(S)$ dont le but est le contrôle moteur, mais au contraire d'avoir un apprentissage dont le but est d'apprendre les conséquences sensorielles des commandes motrices, en s'intéressant à l'espace tangent $\{dS\}$ en un point $S_0 = \psi(M_0, E_0)$. Remarquant que la robustesse dépendra du fait que $\psi$ soit localement linéaire, les auteurs identifient deux sous-espaces distincts : le premier est le sous-espace $\{dS\}_{dE=0}$ des variations sensorielles due aux seuls commandes motrices, et le sous-espace $\{dS\}_{dM=0}$ des variations sensorielles due aux seuls changements de l'environnement. Ils remarquent de plus que $\{dS\} = \{dS\}_{dM=0} + \{dS\}_{dE=0}$.

**Principe de compensabilité :** Le principe à l'œuvre dans ces apprentissages est celui de compensabilité. Une compensation apparaît lorsque la variation sensorielle provenant de $dM$ et celle provenant de $dE$ s'annule. Les variations $dM$ et $dS$ sont alors perçues comme compensables. Cette propriété, qui s'apprend par l'expérience sensorimotrice, mène à percevoir le monde extérieur à travers ses propres mouvements : suivant les intuitions de H.Poincaré, l'idée est justement que cette apprentissage permettent de découvrir et de caractériser le monde extérieur. Ainsi, de nombreux travaux ont porté sur la manière d'inférer la dimensionnalité du monde extérieur, en se basant sur la dimensionnalité des variations sensorielles compensables (voir [Philipona et al., 2003; Laflaquière, 2013]).

**Interaction réelle :** Comme nous l'avons déjà souligné, cet apprentissage a besoin de l'interaction réelle avec le monde. Cette interaction est saisie par les lois de la présence sensorielle réelle : "bodiliness" (capacité à changer les données sensorielles à l'aide de son corps), "insubordinateness" (les entrées sensorielles peuvent également varier sans l'aide du corps), "grabbiness" (les sensations, par des transitions soudaines par exemple, peuvent attirer l'attention). Ces lois peuvent être apprises, et constituent, en quelque sorte, les "marqueurs" du réel.

**Le contexte détermine l'applicabilité :** Le point précédent vaut pour l'apprentissage, toutefois notre cerveau est à même de détecter l'applicabilité d'une loi déjà apprise indépendamment de son application réelle courante (c'est-à-dire à travers la boucle d'interaction sensorimotrice, et donc l'action réelle). Par exemple, lorsque la main est prête à presser une éponge et que la loi relative à la pression de cette éponge a été apprise, son applicabilité peut être détectée [O'Regan and Noë, 2001].

4 Le quatrième niveau d'abstraction consiste en l'accès aux lois sensorimotrices ("cognizing", voir [O'Regan and Block, 2012]). Cet accès cognitif détermine si une personne est consciente d'une expérience. Dans le point précédent nous avons vu que le phénomène de perception est caractérisé par la loi sensorimotrice qui s'applique, que l'interaction soit

réelle ou pas (comme dans le cas du rêve). Nous ajoutons ici que la conscience de cette perception est donnée par l'accès cognitif à cette perception, et donc aux lois et catégories qui s'y rapportent. Notons qu'il y a deux sauts, ou écarts, possible entre la conscience et les lois réelles (physiques) qui s'appliquent. Le premier provient du fait que la conscience est *un certain* accès aux lois qui s'appliquent : cet accès fournit un ensemble de lois potentiellement différent de celui qui se trouvent être effectivement appliqué (et notamment, cet accès peut être "trompé", ce qui constitue la seconde manière d'expliquer le phénomène des rêves [O'Regan and Block, 2012], la première étant citée au niveau 2). Le second écart possible provient du fait que les lois effectivement appliquées sont elle-mêmes potentiellement différentes de celles qui devraient l'être au vu de l'interaction réelle avec l'environnement, de part l'existence du niveau 2. Toutefois, il faut souligner que si les lois accédées par la conscience peuvent être celles de la pression d'une éponge (perception de la mollesse), elles peuvent aussi être celles relatives aux marqueurs du réel ("grabiness" par exemple). De la sorte, si la conscience d'une perception est indépendante de la réalité de cette perception (pourtant à la source des qualia qui y sont associés), celle-ci peut avoir accès au fait que cette perception n'est pas réelle.

En conséquence de cette définition, nous proposons de qualifier d'activité consciente une activité où l'on accède aux lois de cette activité, et ainsi au fait d'*éprouver* un ensemble de lois sensorimotrices identifiées.

De manière parallèle à ces questionnements, il a également été développé des théories concernant un codage commun des senseurs et des moteurs, en observant la relation entre ceux-ci moins du point de vue des lois qui régissent leurs relations ou des perceptions qu'elles entraînent, qu'en considérant comment senseurs et moteurs se lient dans la perception d'un but.

### 1.1.3 Principe idéomoteur

Le principe idéomoteur [James, 1890] postule que les buts sont perçus comme la représentation anticipée des conséquences de l'action, et que les actions volontaires (et la planification de l'action) sont contrôlées par la représentation anticipée des effets désirés. L'idée d'encoder l'action

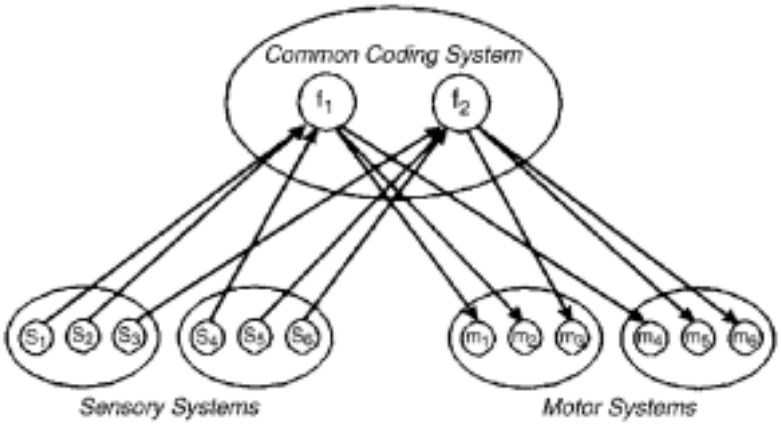


**FIGURE 1.3 –** Illustration de la théorie d'un codage commun entre plusieurs modalités de senseurs et de moteurs. Issu de [Hommel et al., 2001].

en terme de buts a notamment été développée avec la théorie du codage commun ("common coding"), qui propose que les actions soient codées en terme d'effets perceptifs produits [Prinz, 1990]. Dans la théorie de l'"event coding" (TEC) [Hommel et al., 2001], il est précisé que percevoir le stimulus d'un objet, et planifier une action volontaire, ne sont pas des processus distincts opérant sur des encodages différents mais sont fonctionnellement équivalents. Car ils ont tous les deux à faire à une représentation interne des événements externes. Dans la figure 1.3 cette représentation interne est illustrée, et on y voit que des stimuli provenant de différentes modalités (visuelle et auditive dans l'exemple des auteurs) sont codés ensembles, et utilisés pour activer à leur tour différents moteurs (comme la main ou la voix). Ajoutons que Hommel défend l'idée qu'un tel encodage est hiérarchique, et repose sur des événements qui peuvent être segmentés en plusieurs unités ayant chacune du sens.

Le lien entre perception, et planification en vue d'un but, entraîne des questionnements quant à l'influence de la capacité d'un enfant à effectuer une tâche sur sa perception d'une autre personne effectuant cette même tâche. Ainsi, dans [Sommerville and Woodward, 2005] Sommerville et Woodward étudient comment la capacité d'un enfant de 10 mois à résoudre une tâche de "means-end" (voir section 1.1.1) influence sa perception d'une autre personne effectuant cette même tâche, tâche qui est alors perçue comme orientée par un but [Longo and Bertenthal, 2006]. D'autres expériences montrent l'importance de la production de l'action pour la perception de l'action, notamment celles menées par [Hauf et al., 2007] (se référer à [Aschersleben, 2006] pour une étude sur le sujet, et à [Daum et al., 2008] pour une discussion sur ces résultats). Notons qu'il y figure des résultats allant dans ce sens mais avec l'utilisation d'outils, dans lesquels la seule observation de l'utilisation d'outils a des effets proches de ceux décrits dans la section 1.1.1 concernant l'espace atteignable.

Il faut noter que considérer les actions comme provenant d'une représentation anticipée des effets désirés, comme c'est le cas pour le principe idéomoteur, interroge l'action d'une manière différente que ne le font les associations stimulus-réponses, ou stimulus-récompenses.

Pour le principe idéomoteur la représentation des effets désirés implique immédiatement une planification d'actions. Pourtant l'émergence de séquences d'actions soulève des problématiques propres, et nous allons les aborder dans la partie qui suit.

### 1.1.4 Séquence d'actions

La planification de l'action soulève d'importantes questions concernant la manière dont une séquence d'actions unitaires significatives peut être organisée. Des travaux fondateurs [Lashley, 1951; Miller G. A., 1960] proposent une structure hiérarchique de l'action humaine. Allant dans ce sens, un modèle computationnel de schémas d'actions, organisé hiérarchiquement, et destiné à être associé à la structuration propre du domaine d'une tâche donnée (par exemple, faire le café) a été proposé par Cooper et Shallice [Cooper and Shallice, 2000] (voir figure 1.4, et [Botvinick, 2008] pour une revue).

Mais on retrouve également en neuroscience ([Dehaene and Changeux, 1997; Chersi et al., 2011; Grossberg, 1987]) ainsi qu'en robotique ([Nicolescu and Matarić, 2002; Ogino et al., 2012; Sandamirskaya and Schöner, 2010; Billing and Sandamirskaya, 2015]) des architectures pour la planification hiérarchique, et l'organisation hiérarchique du comportement. Ainsi, pour Bonini *et al.* les neurones du cortex prémoteur ventral et ceux du pariétal inférieur, avec ceux du cortex préfrontal, peuvent encoder le but d'actions intentionnelles à différents niveaux d'abs-

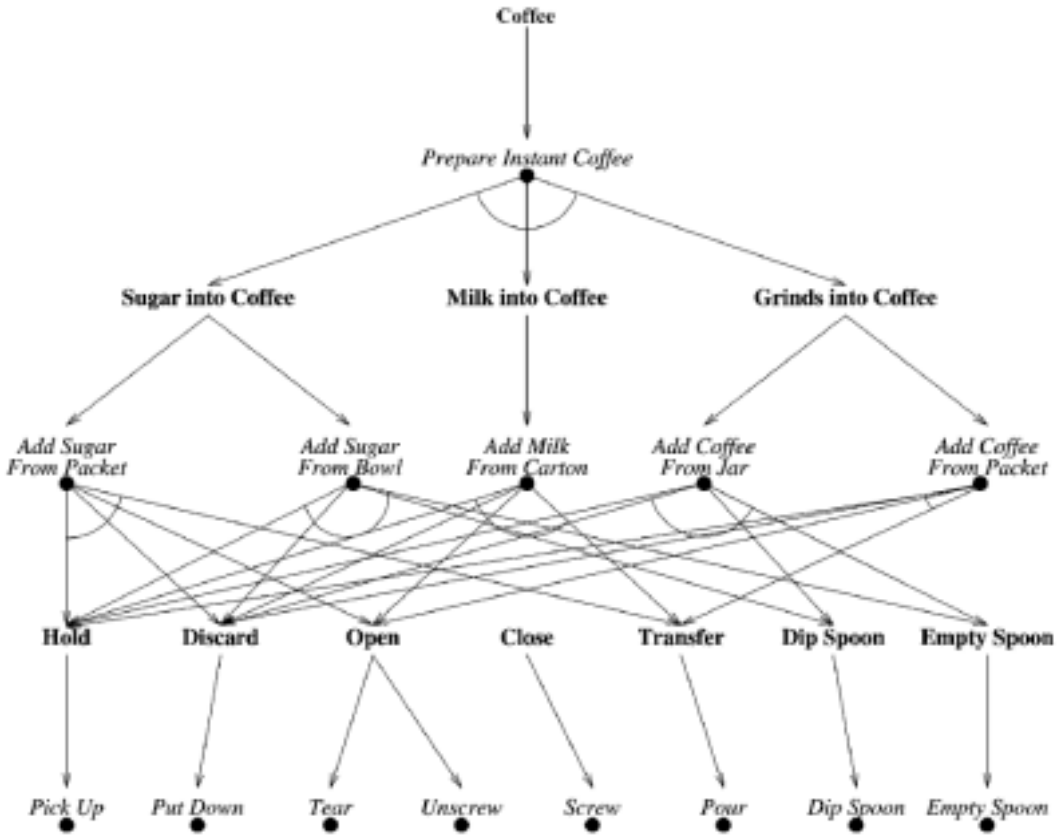


**FIGURE 1.4 –** Organisation des schémas (actions, en italique) et buts (en gras) dans le domaine de la préparation du café. Issu de [Cooper and Shallice, 2000]

traction motrice. De plus, le circuit pariéto-prémoteur travaille avec le cortex préfrontal pour organiser les actions motrices en séquence d'actions, et garder actif la représentation interne de l'intention motrice [Bonini et al., 2011].

D'un point de vue différent, la théorie des systèmes dynamiques montrent que ces séquences et cette organisation hiérarchique peuvent être le résultat de bifurcations dans un système dynamique plus global.

Dans [Billing et al., 2015], des champs de neurones dynamiques (DNF, voir [Schöner, 2008]) sont proposés pour combler le fossé entre les dynamiques sensorimotrices bas niveaux et des processus cognitifs plus hauts niveaux, tel que la planification. Les auteurs proposent des comportements élémentaires tels que "chercher un objet de la couleur X" ou "bouger le bras à la position Y" associés avec des DNF conditionnant l'initialisation et la finalisation (conditions de satisfaction) de ces comportements. Chaque comportement est aussi associé avec un champ de motivation (permettant de commencer le comportement) et un champ de précondition (pour savoir si le système est à même de le faire). Une fois que le champ de motivation est activé, le DNF gérant l'initialisation est aussi activé mais peut être inhibé si le champ de précondition ne satisfait pas les conditions nécessaires. Dans ce cas, ce champ peut activer le champ de motivation d'un autre comportement élémentaire qui permettra de remplir cette condition. Une fois

celle-ci effectivement satisfaite, le champ de précondition arrête l'inhibition du comportement (voir fig. 1.5).

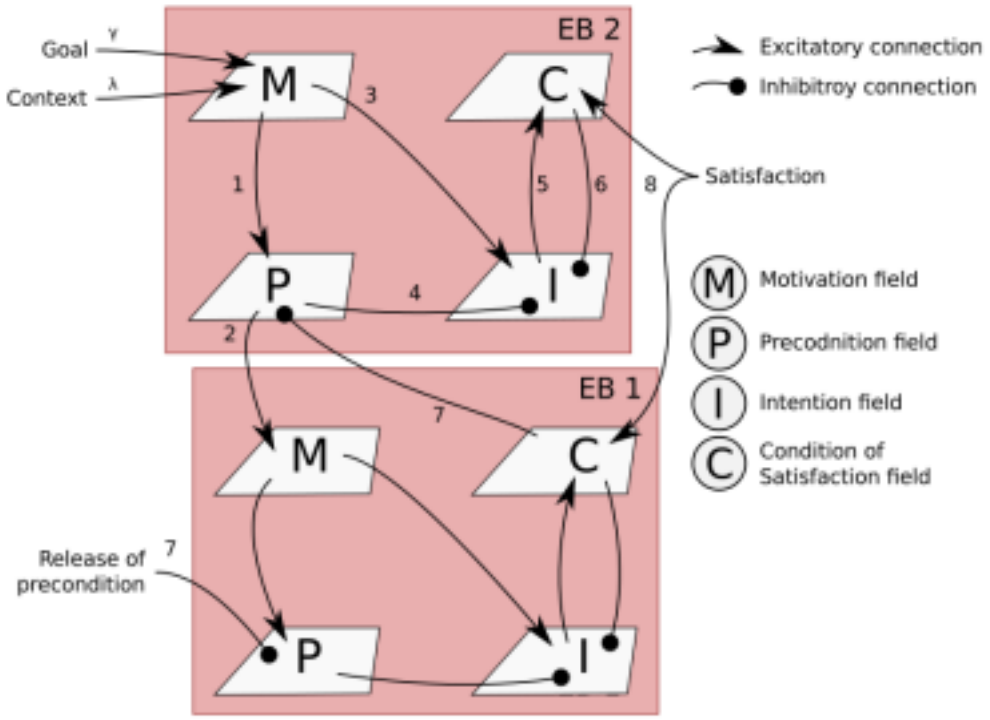


**FIGURE 1.5 –** Une séquence de deux comportements élémentaires EB1 et EB2. Un état but est ici représenté comme une motivation M à exécuter EB2, en ayant satisfait les préconditions P. Issu de [Billing et al., 2015].

Comme nous l'avons souligné en section 1.1.1, la capacité à faire des séquences, si elle est importante, n'est pas suffisante pour utiliser des outils. La théorie des affordances semble essentielle pour comprendre cette aptitude, et nous allons nous y intéresser dans la section qui suit.

### 1.1.5 Affordances, neurones miroirs et théorie de la simulation

#### Neurones miroirs

Les neurones miroirs ont été découvert sur les macaques, dans la zone F5 du cerveau (considérée comme la partie analogue du cortex frontal inférieur postérieur de l'humain). Ces neurones miroirs répondent lorsque des actions spécifiques sont effectuées, comme par exemple attraper un objet, mais aussi lors de l'observation d'un autre macaque ou d'un humain effectuant la même action, lorsque celle-ci est orientée par le même but. Plus précisément, ces neurones vont répondre différemment si cette même action (attraper une chose) est observée au sein de séquences motrices différentes, et est effectuée avec une intention différente (attraper cette chose *pour* la manger, ou la déplacer par exemple). Ceci semble indiquer un mécanisme en lien avec le but de l'action (voir [Rizzolatti and Craighero, 2004; Craighero et al., 2007] pour une re-

vue). Pour [Fogassi et al., 2005], les résultats suggèrent que ces neurones sont sensibles au but poursuivi tant au niveau proximal, que distal.

Notons que, en allant plus loin, pour Rolf et Asada les buts peuvent souvent être compris en terme d'affordances sous-jacentes : les buts traduisent alors un écart entre l'état présent et un état désiré imaginé [Rolf and Asada, 2015]. Les affordances décrivent dans ce cas des possibilités générales, qui peuvent être valuées par des buts.

### Affordances

Les neurones canoniques furent découverts dans la même aire (zone F5) du cerveau. Ceux-ci répondent aussi à des actions spécifiques, comme attraper un objet, mais également lors de la seule observation de ce même objet [Rizzolatti and Craighero, 2004]. Ceci semble indiquer un mécanisme d'affordance. Le concept d'affordance fut initialement introduit par Gibson [Gibson, 1979] en tant que ressources que l'environnement offre à tout animal doté de la capacité de les percevoir et de les utiliser. Notons que du point de vue de cette perspective, écologique, la perception des affordances est directe et ne solicite pas la médiation de représentations neuronales ou mentales. Du point de vue de la perspective représentationnelle toutefois, les affordances sont internalisées par des représentations mentales afin d'être utilisées par les processus computationnels, par exemple avec un agent sensorimoteur interagissant avec le monde.

Les outils, par exemple, sont des objets manipulables générant de multiples affordances, notamment en relation avec leurs formes ou leurs fonctions [Creem-Regehr and Lee, 2005; Bub

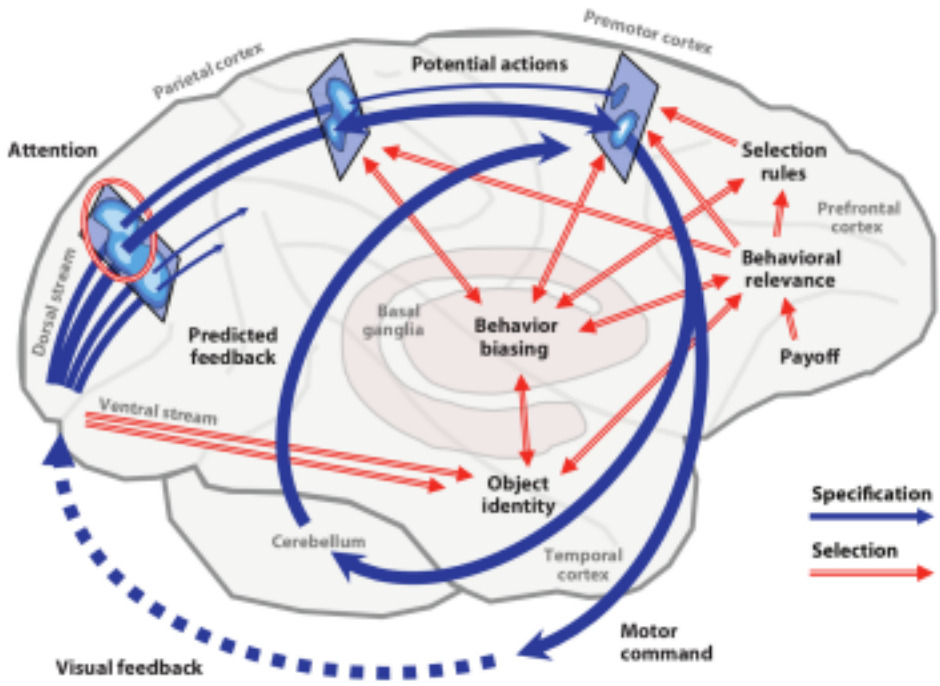


**FIGURE 1.6 –** Hypothèse de la compétition entre affordances, dans le contexte de mouvements guidés par la vision. Les flèches bleues et pleines illustrent les voies dédiées à la spécification de l'action, également utilisées pour guider son exécution, tandis que les flèches rouges illustrent les voies permettant la sélection de l'action. Issu de [Cisek and Kalaska, 2010].

et al., 2008]. Ellis et Tucker ont proposé le concept de micro-affordances, pour des interactions non avec l'objet complet mais avec des composantes particulières permettant l'action. Par exemple pour une taille d'objet donnée, ou son orientation particulière, il sera déclenché l'activation partielle des circuits moteurs requis pour interagir avec celui-ci [Ellis and Tucker, 2000].

### Compétition entre affordances

Puisque l'environnement fourni en permanence au cerveau de nombreuse opportunités d'actions, en parallèle, cela pose le problème de leur interférence, et donc, plus généralement, d'un mécanisme de sélection.

Selon l'hypothèse de la compétition des affordances, développée par Cisek [Cisek, 2007; Cisek and Kalaska, 2010], les questions relatives à ce que l'on doit faire (sélection de l'action) et comment on doit le faire (spécification de l'action) sont traitées non successivement mais en parallèle. Le cerveau traite donc l'information sensorielle et spécifie, en parallèle, plusieurs actions potentielles disponibles, les affordances. Parallèlement, une autre voie du cerveau est supposée biaiser la sélection parmi les affordances disponibles, lesquelles sont en compétition les unes avec les autres (voir fig. 1.6). Dans de récents travaux, Thill *et al.* ont publié une étude portant sur ces mécanismes (et les modèles computationnels qui y sont liés) en insistant sur le rôle des buts, et plus spécifiquement du cortex préfrontal dans cette sélection [Thill et al., 2013].

### Affordances et simulations

Les travaux de Norman [Norman, 2002] soulignent que les exemples d'affordances donnés par Gibson, à savoir percevoir le sol comme "marchable", l'eau comme "nageable" ou une chaise comme "asseyable", impliquent à chaque fois une action de l'observateur, comme marcher, nager ou s'asseoir. Poursuivant cette idée, Moller *et al.* supposent que les affordances sont révélées par un processus de simulation interne des conséquences de l'action [Möller and Schenck, 2008]. En se basant sur l'information sensorielle présente, l'observateur va effectuer des simulations internes de séquence d'actions, et anticiper leurs effets. Ainsi, en opposition avec la vision écologiste de Gibson, les affordances seraient générées activement par un processus de simulation sensorimotrice, influencé par les données sensorielles, plutôt que directement obtenues par la seule analyse passive de ces données sensorielles.

Notons qu'afin d'effectuer de telles simulations, une capacité d'anticipation est requise. L'anticipation est un mécanisme fondamental du cerveau, et il est soutenu que l'apprentissage prédictif de l'information sensorimotrice joue un rôle clé dans le développement cognitif [Butz et al., 2003; Barsalou, 2009; Barsalou et al., 2007; Nagai and Asada, 2015]. L'anticipation des effets de l'action est généralement effectuée par des modèles internes (direct et inverse), comme on le verra dans la section 1.2.1.

Ainsi, dans [Möller and Schenck, 2008], les affordances sont issues de la simulation de l'action, en utilisant des modèles directs. Il sera alors possible de prédire et choisir les actions à effectuer en ce basant sur ces simulations. Celles-ci reposent sur des modèles directs (implémentées sous la forme de perceptrons multi-couches) qui prédisent des variations sensorielles basées sur l'état courant et le mouvement effectué (voir fig. 1.7). Pour des raisons de performances, le modèle de la figure 1.7 a été par la suite séparé en autant de prédicteurs différents par

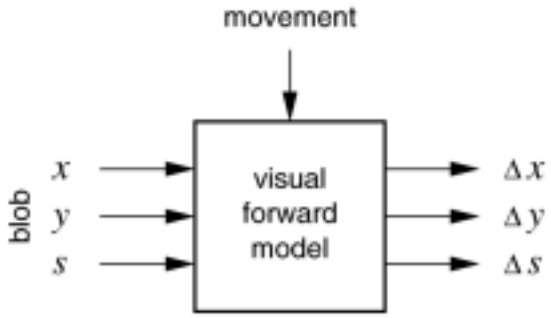


**FIGURE 1.7 –** Modèle direct prédisant la variation sensorielle pour la vision. Issu de [Möller and Schenck, 2008].

senseur dont la variation est prédite, chacun ayant par contre les mêmes entrées (position $x$ et $y$ et la taille $s$ visuelle du blob observé).

L'idée est alors d'utiliser un processus de simulation interne en utilisant le modèle direct : en prenant l'état courant $S(t)$ en entrée, et une entrée motrice simulée $M(t)$, le modèle prédit $S(t+1)$ en sortie, laquelle peut-être utilisée en entrée d'une autre simulation. Pour un modèle prédisant une variation visuelle et tactile, le processus de simulation interne est montré figure 1.8.

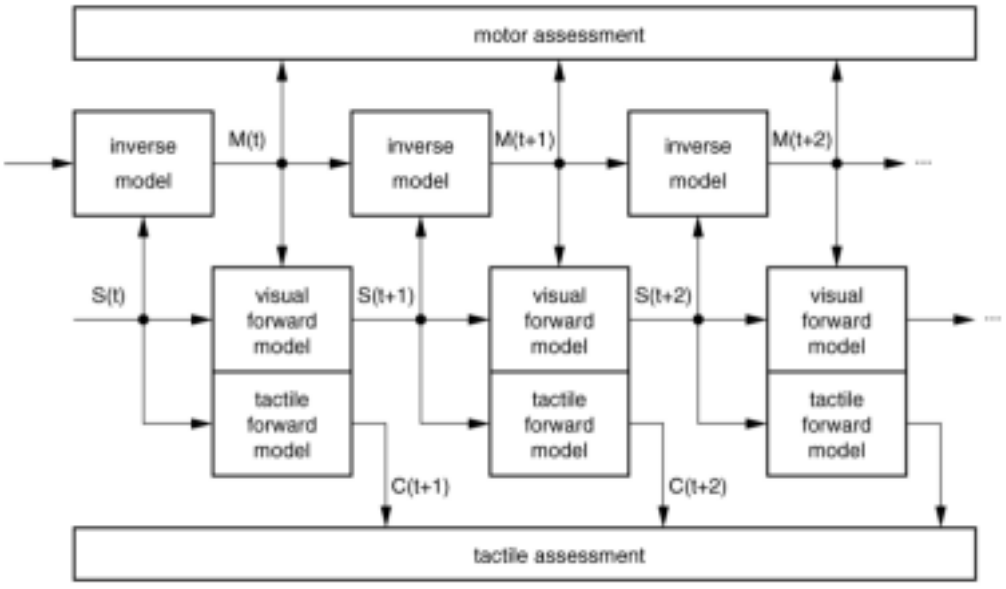


**FIGURE 1.8 –** Processus de simulation interne montrant l'interaction entre modèles direct et inverse. Issu de [Möller and Schenck, 2008].

Cependant, la multitude des séquences possibles étant trop riches pour être explorée, un modèle inverse est appris afin de réduire cette exploration en suggérant des actions. À l'aide d'un mécanisme d'évaluation, les actions sont alors étiquetées comme "bonne" ou "mauvaise", et le modèle inverse est alors à même de fournir un ensemble restreint de séquences prometteuses.

### Contrôle et théorie de la simulation

Dans la théorie de la simulation (ou de l'émulation), il a été défendu que par la simulation mentale, une même structure neuronale pouvait sous-tendre des capacités sensorimotrices, cognitives et sociales aussi varié que la détection de soi, la distinction soi-autrui, la planification, la perception, ou encore l'imitation [Barsalou, 1999; Hesslow, 2002; Grush, 2004; Decety and Grèzes, 2006; Pezzulo et al., 2013a; Schillaci et al., 2016].

Ainsi la théorie de la simulation ne fournit pas seulement des mécanismes susceptibles de proposer une explication aux affordances ([Möller and Schenck, 2008]), mais constitue également un moyen de comprendre le lien bi-directionnel, entre la représentation des buts, et les programmes moteurs permettant de satisfaire ces buts (tel que le suppose le principe idéomoteur et la théorie du common coding).

Pour Pezzulo et Castelfranchi [Pezzulo and Castelfranchi, 2009], penser consiste ainsi à contrôler son imagination : il s'agit de la capacité à contrôler les simulations mentales afin de mettre en place des sous-buts, et planifier au delà de l'ici et maintenant de la perception. Ils proposent, comme dans [Möller and Schenck, 2008] une extraction parallèle des nombreuses affordances disponibles dans l'environnement en faisant courir des simulations internes des actions possibles. Comme dans [Cisek, 2007], ils proposent également un mécanisme pour sélectionner l'une des affordances disponibles en compétition, en se basant sur leur valeur et leur faisabilité. Enfin, ils suggèrent deux mécanismes d'inhibition. L'un permet d'inhiber les nombreuses réponses automatiques, déclenchées par les affordances disponibles. Le second permet à la fois l'inhibition temporaire des commandes motrices, et l'inhibition des entrées extérieures, afin de leur substituer des stimuli générés en interne (simulés). Ils soulignent enfin l'importance d'une mémoire de travail permettant de mémoriser des buts distants ainsi que des sous-buts au cours du temps (voir fig. 1.9).

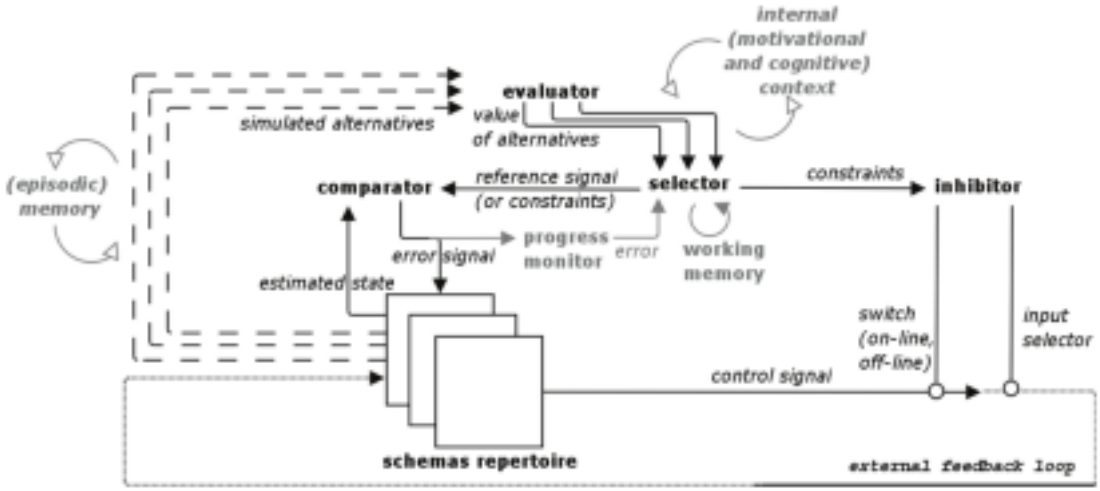


**FIGURE 1.9 –** Modèle théorique présenté par Pezzulo et Castelfranchi, intégrant simulations et compétitions entre affordances dans le cadre d'actions orientées vers un but. Issu de [Pezzulo and Castelfranchi, 2009].

Nous tenons ici à souligner l'importance de la distinction entre comportements orientés par un but, et les comportements habituels, ou de type stimulus-réponse. La psychologie ainsi que

la neurobiologie attestent que les uns ne sauraient se substituer aux autres et que chacun joue un rôle essentiel, irréductible à l'autre. Se pose alors la question, ouverte, de leur articulation. Afin d'illustrer comment de tels mécanismes pourraient cohabiter, nous pouvons noter l'exemple de Pezzulo *et al.* [Pezzulo et al., 2013b] qui proposent un modèle, appelé Contrôleur Instrumental Mixte (MIC), permettant d'équilibrer de manière flexible chacun de ces comportements, en combinant ceux basés sur des modèles et ceux non basés sur des modèles : "model-free" pour les comportements habituels et "model-based" pour les comportements orientés par un but, voir figure 1.10.

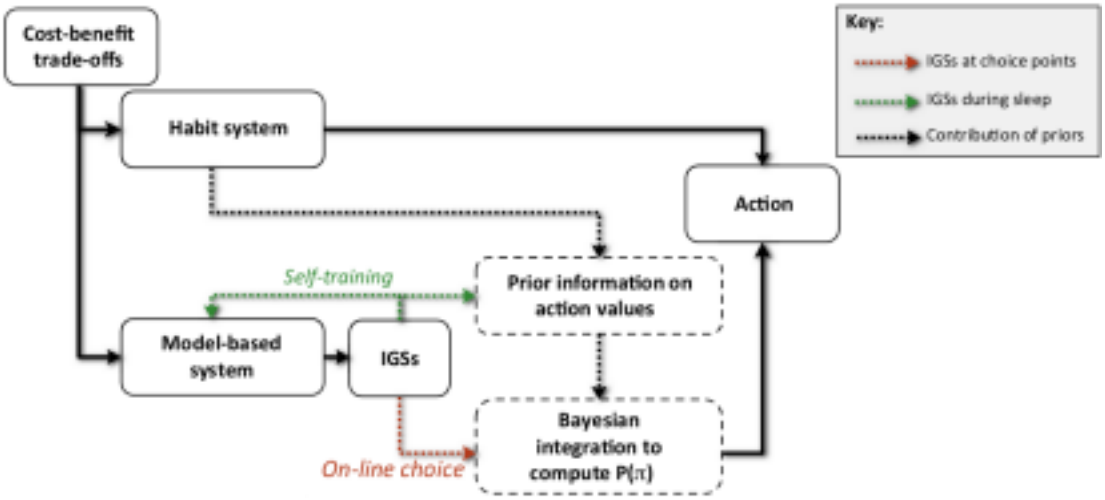


**FIGURE 1.10 –** Contrôleur mixte combinant, en fonction de leur coûts et bénéfices, des comportements habituels et des comportements orientés par un but (et permettant notamment de gérer des séquences - IGS). Issu de [Pezzulo et al., 2014].

Après avoir étudié notre problématique sous l'angle de théories venant des sciences cognitives, de la psychologie développementale ou des neurosciences, dans la partie qui suit nous l'aborderons sous l'angle de la robotique, afin d'y observer les problématiques propres, et les types de solutions qui sont généralement proposées.

## 1.2 Apprentissage en robotique

Dans cette partie, qui ne vise pas à l'exhaustivité, nous commencerons par traiter la question générale de ce que peut être un apprentissage sensorimoteur en robotique (Sec. 1.2.1). Puis nous nous pencherons sur les approches les plus répandues relatives au contrôle en robotique, à savoir le contrôle optimal (Sec. 1.2.2), ainsi que l'apprentissage par renforcement (Sec. 1.2.3). À la suite d'une analyse des limitations concernant l'existence de fonctions à optimiser, nous nous intéresserons au travaux précédents du laboratoire en considérant l'approche PerAc (Sec. 1.2.4) qui s'attache aux propriétés émergentes de l'apprentissage de la boucle perception-action. Nous proposerons alors un aperçu de la manière d'aborder la question spécifique de l'utilisation d'outils en robotique (Sec. 1.2.5). Enfin, nous regrouperons les propriétés que divers auteurs ont retenu comme essentielles pour la possibilité d'un apprentissage autonome et "open-ended" (Sec. 1.2.6).

### 1.2.1 Apprentissage sensorimoteur

En robotique, la problématique de l'encodage de l'information sensorimotrice peut être illustrée par la manière d'effectuer l'encodage moteur, couplé à celui de l'effecteur terminal d'un bras robotique (la main robotique). C'est un problème que l'on retrouve notamment lors de tâches consistant à atteindre un point donné ("reaching") pour, par exemple, attraper quelque chose avec la main.

Les roboticiens sont confrontés au problème de l'équivalence motrice, ou de la redondance. Ce problème apparaît pour tout effecteur doté de plus de dimensions que celles dans lesquelles sont spécifiées une cible. Ainsi, pour un bras disposant de suffisamment de degrés de liberté, il existe une infinité de trajectoires motrices permettant d'atteindre un point donné.

Pour Bullock et Grossberg [Bullock et al., 1993], il existe deux stratégies principales permettant l'encodage sensorimoteur liant les coordonnées motrices à celles de la main robotique, permettant d'effectuer des tâches de "reaching" (voir la fig. 1.11, et [Bullock et al., 1993; Braud et al., 2015] pour plus de détails).

La première stratégie consiste à apprendre la correspondance ("mapping") entre ces deux systèmes de coordonnées. Un robot peut ainsi atteindre une cible donnée $X_B$ en activant les coordonnées motrices $\theta_B$ associées avec la position désirée de la main.

$$\theta_B = f(X_B) \tag{1.1}$$

Nous appelons cette stratégie "absolue" parce qu'une correspondance absolue est apprise entre chaque coordonnée (voir la fig. 1.11).

Les algorithmes basés sur cette stratégie peuvent par exemple considérer cet apprentissage comme étant celui d'un système homéostatique : après avoir appris ces correspondances, le système, indépendamment du bruit ou d'un quelconque changement, sera forcé d'équilibrer ses sorties sur un point d'équilibre basé sur les catégories apprises, en cas de conflit entre ses entrées visuelles et proprioceptives. Dans cette situation, les catégories apprenant cette correspondance peuvent être considérées comme des attracteurs visio-moteurs [Gaussier and Zrehen, 1995; De Rengervé et al., 2015]. Si le robot doit effectuer une séquence continue avec la main dans l'espace visuel (par exemple pour former un "8"), le robot générera alors une séquence

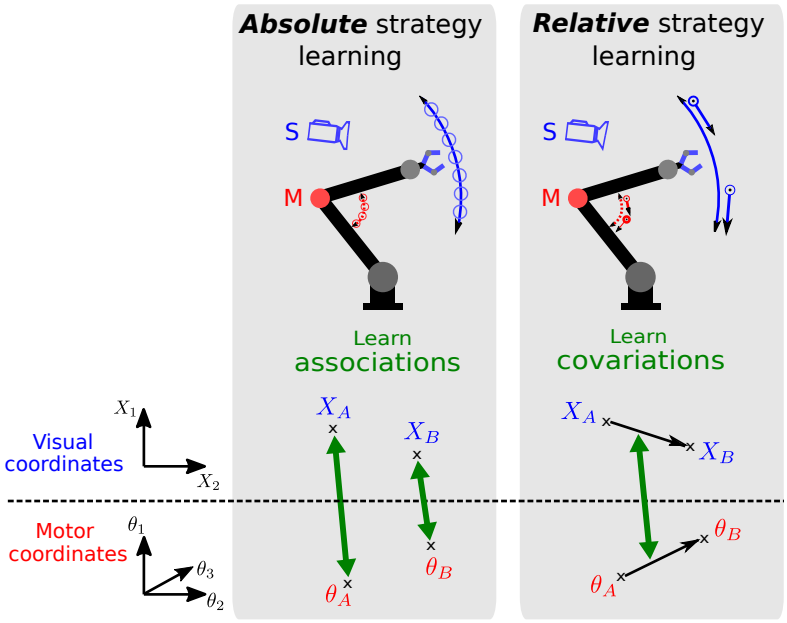


**FIGURE 1.11 –** Stratégie absolue et relative pour l'encodage sensorimoteur

continue de configurations motrices permettant d'effectuer cette tâche. Notons qu'avec cette stratégie, un algorithme dédié à la minimisation de la distance entre les positions courantes et désirées de la main est requis (comme un simple PID, par exemple).

La seconde stratégie, appelée "relative", consiste à apprendre la correspondance entre chaque direction spatiale constatée par la main, et le changement correspondant dans les angles des articulations motrices qui ont provoqué ce mouvement (voir la fig. 1.11). Cette stratégie a pu faire l'objet d'implémentations neuronale, inspirées de données neurobiologiques (voir [Burnod et al., 1992, 2000]). Mais généralement, cette approche implique l'utilisation d'un modèle direct à travers la matrice jacobienne $J$ :

$$\dot{X} = J(\theta)\dot{\theta} \tag{1.2}$$

Avec cette stratégie, un robot peut atteindre une cible non de manière immédiate, comme dans le cas de la stratégie absolue, mais à travers la direction allant de la main $X_A$ à la cible $X_B$. Cette direction sera alors associée avec la variation correspondante de l'angle des articulations motrices, laquelle pourra alors être effectuée pour atteindre la cible.

$$\dot{\theta}_B = f(X_B - X_A) \tag{1.3}$$

La fonction permettant cela est généralement estimée grâce à des algorithmes d'apprentissage inverse dédiés. Généralement, les méthodes consistent à inverser la matrice jacobienne [Baillieul et al., 1984; Mussa-Ivaldi and Hogan, 1991; D'Souza et al., 2001] afin de sélectionner une unique solution, parmi la multitude des possibilités offertes, qui permettra d'effectuer la tâche.

$$\dot{\theta} = J^{+}(\theta)\dot{X} \tag{1.4}$$

Les modèles inverses sont bien souvent accompagnés de prédicteurs permettant d'anticiper, appelés modèles directs (voir Fig.1.12). Le modèle direct requiert une intégration des états perceptuels ainsi qu'une copie efférente des commandes motrices afin de prédire les conséquences sensorielles des actions. Le modèle inverse requiert, quant à lui, l'état actuel et désiré afin de générer les commandes motrices permettant d'atteindre l'état désiré (voir [Wolpert et al., 1998; Kawato, 1999; Wolpert and Ghahramani, 2000; Butz et al., 2003, 2006]).

Se fondant sur l'idée que le monde est modulaire, Wolpert et Kawato ont proposé un modèle basé sur des paires de modèles directs et inverses [Wolpert and Kawato, 1998], dans lesquelles les modèles directs apprennent la relation causale entre les actions et leurs conséquences, et sont ensuite utilisés comme prédicteur ou simulateur des conséquences des actions [Wolpert et al., 2003] (voir fig.1.13 extrait de [Wolpert et al., 1998]).

Cependant, une représentation modulaire du monde tient lorsqu'il y a une indépendance entre les modules, ce qui peut être notamment le cas pour des systèmes linéaires, mais le monde est tel que ses différentes parties sont en interaction dynamique non linéaire. Or définir ces modules, et leurs frontières, définit également les comportements qui leur sont adaptés, et la manière d'y résoudre des tâches. Ceci peut donc entraîner des limitations pour certaines tâches, et peut rendre difficile le passage du mono au multi-tâches.

Si les questions concernant l'encodage sensorimoteur sont centrales pour notre problématique, tant du point de vue de la psychologie (voir sections 1.1.2 et 1.1.3) que du point de vue robotique, le seul fait de travailler au contrôle d'un bras robotique afin qu'il bouge et puisse

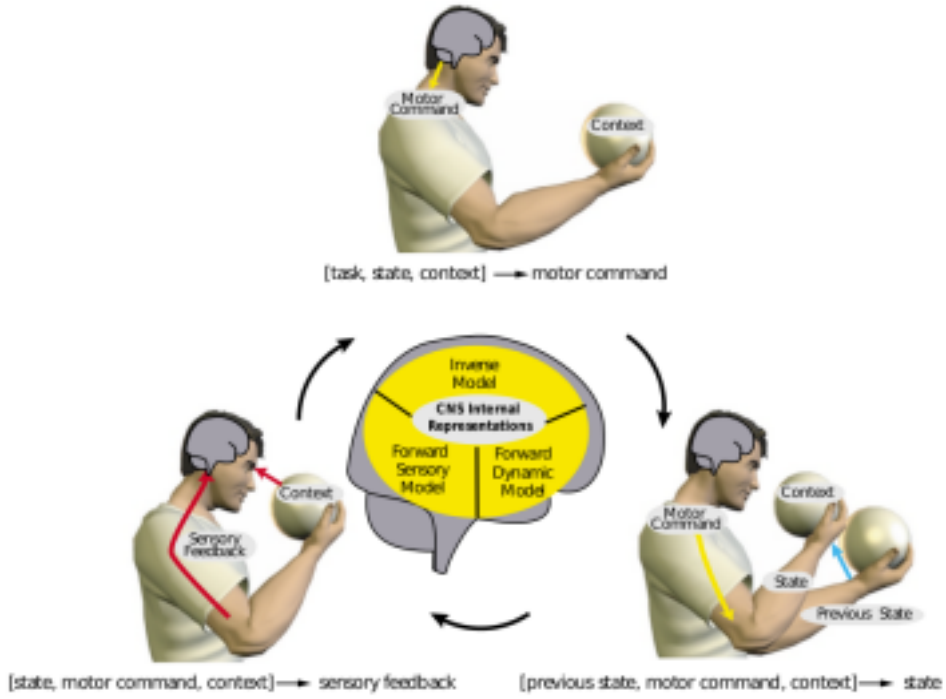


**FIGURE 1.12 –** Boucle sensorimotrice montrant la génération de commandes motrices proposées par le modèle inverse en haut, et en bas les modèles directs prédisant : les transitions d'états (à droite), et les retours sensoriels (à gauche) au sein du système nerveux central (CNS). Issu de [Wolpert and Ghahramani, 2000]

éventuellement effectuer une tâche, quelle qu'elle soit, soulève des questions quant à la manière d'effectuer ce contrôle et à ce qu'on attend de lui. Dans la partie qui suit, nous étudierons des approches classiques permettant d'effectuer le contrôle d'un robot pour effectuer une tâche.

### 1.2.2 Théorie du contrôle optimal

Dans une approche que l'on peut considérer "classique" de la robotique, l'intérêt est principalement porté sur l'accomplissement de tâches, et ce sera donc sur la manière de l'accomplir que les évaluations porteront. Ainsi, la distance à la cible, la trajectoire empruntée ou encore le temps mis à accomplir certaines tâches, étant évaluables, ont donné lieu à diverses propositions de fonctions de coûts (voir Sec. 1.2.2). L'enjeu est alors l'optimisation du comportement des robots en fonction de ces fonctions de coûts, et la théorie qui le décrit est celle du contrôle optimal. Si le cadre d'application peut être déterministe, il est également possible de l'inscrire dans le cadre stochastique, plus générique encore, et ainsi être appliqué aux Processus Décisionnel de Markov (MDP). (Voir une revue des questions y sont liées dans [Sigaud and Buffet, 2008], dont nous nous sommes ici inspirés). L'hypothèse de Markov stipule que la distribution de probabilité spécifiant un état donné ne dépend que de l'état précédent et de l'action effectué, mais pas du passé. De manière analogue, le principe d'optimalité stipule que le choix de l'action optimale dans le futur est indépendant des actions qui ont précédé, et qui ont mené à l'état présent.

Un état est ce qui résume la situation de l'agent à chaque instant, et une action (ou décision) influence la dynamique de l'état. Il faut ajouter la notion de revenu (ou récompense) qui est associée à chacune des transitions d'état. Les MDP sont des chaînes de Markov qui visitent ces états, contrôlées par les actions et valuées par les revenus. Résoudre un MDP revient à contrôler

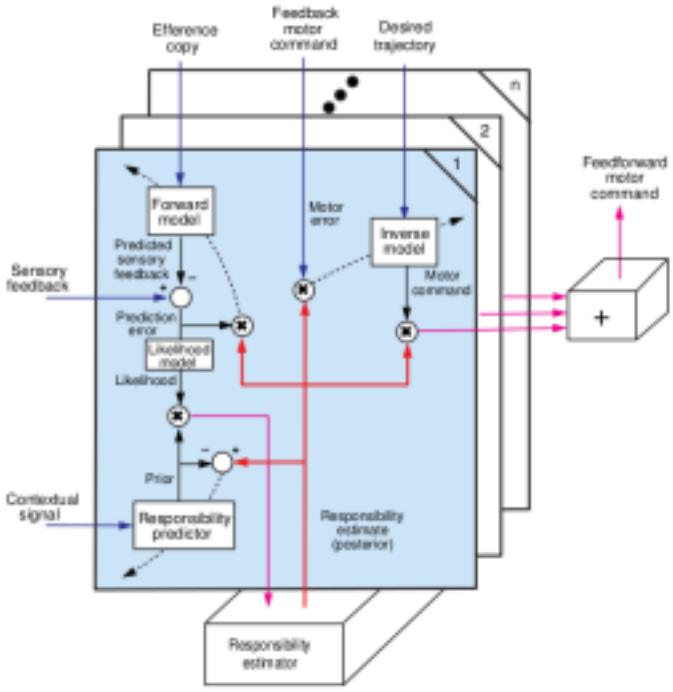


**FIGURE 1.13 –** Modèle de Wolpert et Kawato extrait de [Wolpert et al., 1998]

l'agent pour optimiser son comportement, c'est-à-dire en maximisant son revenu. Une solution à un MDP s'appelle une politique.

On appelle :

- $S$ l'espace d'états dans lequel évolue le processus
- $A$ l'espace des actions qui contrôlent la dynamique de l'état
- $T : S\text{x}A\text{x}S \to [0, 1]$ sont les probabilités de transition entre états. $T(s, a, s') = p(s'|a, s)$ est la probabilité d'atteindre l'état $s'$ à partir de l'état $s$ après exécution de l'action $a$
- $R : S\text{x}A \to \mathbb{R}$ la fonction de récompense sur les transitions entre états
- $\pi : S\text{x}A \to [0, 1]$ une politique qui code la façon dont l'agent va se comporter (la probabilité de faire une action dans un état donné).

**Programmation dynamique**

Une politique optimale vise à obtenir une séquence de récompenses la plus importante possible, autrement dit à maximiser le cumul des récompenses instantanées le long d'une trajectoire. Ce cumul est donné (ou estimé) à travers une fonction de valeur $V$ :

$$\forall \pi, V^\pi : S \to \mathbb{R} \tag{1.5}$$

Par exemple, dans le cas déterministe, et en supposant $N$ étapes, le critère fini en partant de l'état $s$ et suivant la politique $\pi$ donne la fonction de valeur suivante :

$$\forall s \in S, V_N^\pi(s) = \sum_{t=0}^{N-1} r_t \tag{1.6}$$

Le choix de la politique $\pi$ s'effectue donc ainsi :

$$\pi^* = \arg\max_\pi(V^\pi) \tag{1.7}$$

Toujours dans le cas déterministe, en notant $suivant(s, a)$ l'état dans lequel se retrouve le système en réalisant $a$ à partir de $s$, on définit la politique optimale $\pi$ par

$$\pi^*(s) = \arg\max_{a \in A}(r(s, a) + V(suivant(s, a))) \tag{1.8}$$

Ce maximum est susceptible d'être atteint pour plusieurs valeurs de $a \in A$, et la politique peut ne pas être unique. La fonction de valeur optimale $V^*$ est, elle, unique pour l'état $s$ et vérifie :

$$V^*(s) = \max_\pi(V^\pi) \tag{1.9}$$

$$= \max_{a \in A}(r(s, a) + V(suivant(s, a))) \tag{1.10}$$

Les équations 1.8 et 1.10 sont les équations de Bellman. Notons que ce principe a été étendu au cas continu. Les équations utilisées sont alors les équations de Hamilton-Jacobi-Bellman.

Notons également que la théorie du contrôle optimal est basée sur deux principes fondamentaux : le premier est celui de la programmation dynamique [Bellman, 1957], et le second est le principe du maximum de Pontryagin [Pontryagin et al., 1962]. Ce dernier permet de donner une condition nécessaire d'optimalité tout en évitant le problème de la malédiction de la dimension auquel la programmation dynamique est sujette [Bellman, 1957].

Toutefois, le principe du maximum de Pontryagin ne s'applique qu'au cas déterministe. Le principe d'optimalité de la programmation dynamique de son côté peut être appliqué au cas stochastique des MDP, en tenant compte d'une loi de probabilité $p$ sur l'état d'arrivée $s$ quand on réalise l'action $a$ dans l'état de départ $s_0$ : $s = p(s|s_0, a)$. Dans ce cas, on adaptera les équations de Bellman en prenant l'espérance de la fonction de valeur. Par exemple, le cumul espéré à l'aide du critère fini dans le cas stochastique devient, avec $E^\pi$ l'espérance mathématique sur l'ensemble des réalisations du MDP en suivant la politique $\pi$ :

$$\forall s \in S, V_N^\pi(s) = E^\pi[\sum_{t=0}^{N-1} r_t | s_0 = s] \tag{1.11}$$

Pour finir, nous soulignerons donc que la politique choisie dépend à la fois :

- de la fonction de valeur $V$ utilisée, et donc du critère (fini, $\gamma$- pondéré, total, moyen, *etc.*) utilisé pour cumuler l'espérance des récompenses
- de la fonction de récompenses utilisée (ou de la fonction de coût).

#### Intérêts et limitations de l'optimisation d'une fonction de coût

Un coût est une fonction scalaire dépendant de l'état courant du robot et de ses senseurs. Il s'agit généralement de minimiser la consommation d'énergie, de lisser la trajectoire ou encore d'optimiser la précision obtenue sur le mouvement dans le cas stochastique. Les modèles classiques de contrôle optimal en robotique s'appuient sur la minimisation de l'une de ces caractéristiques pour générer une trajectoire désirée en boucle ouverte (voir [Todorov, 2004] pour plus de détails).

Parmi ces modèles, on trouve ainsi ceux dont le but est de réduire le coût énergétique pour les muscles [Hatze and Buys, 1977]. Différents types de calcul peuvent alors être utilisés : celui-ci peut prendre en compte le couple [Nelson, 1983], considérer un modèle précis du muscle [Taniai and Nishii, 2007; Nishii and Taniai, 2009] ou encore chercher à minimiser les périodes de co-activation musculaire, selon le principe d'inactivation proposé dans [Berret et al., 2008; Gauthier et al., 2010].

Cependant la seule minimisation de cette énergie ne permet pas de rendre compte de la moyenne des comportements relatifs aux mouvements des bras ou des saccades occulaires, ou même de certains mouvements concernant le corps entier.

Ainsi d'autres critères peuvent être minimisés comme les secousses (*i.e.* la dérivée de l'accélération de la main) (minimum-jerk, [Hogan, 1984; Flash and Hogan, 1985]) et la dérivée du couple (minimum torque change, [Uno et al., 1989; Nakano et al., 1999]). Ce type de modèle permet d'obtenir de meilleurs résultats concernant la reproduction des dynamiques de mouvements d'atteintes constatés chez les humains, et permet également de mieux rendre compte de la relation vitesse-curvature que la loi de puissance [Lacquaniti et al., 1983].

Afin d'éviter de ne reproduire que la moyenne des résultats désirés, tout en ayant parfois des erreurs individuelles substantielles, la minimisation de la variance a également été considéré, en minimisant par exemple la variance sur la position de l'extrémité du bras à la fin d'un geste [Harris and Wolpert, 1998]. De plus, des études sur la modélisation du contrôle optimal ont également montré que la loi de Fitts (selon laquelle le temps nécessaire pour atteindre une cible dépend tant de la distance, que de la taille de la cible), peut émerger de la présence d'un bruit dont la variance augmente proportionnellement à la magnitude de la commande motrice envoyée [Tanaka et al., 2006; Guigon et al., 2008].

Notons que le coût le plus pertinent à utiliser, pour un système sensorimoteur donné, peut ne pas correspondre à notre compréhension intuitive de la tâche que nous voulons voir ce système réaliser. Par exemple, s'il semble évident que le système nerveux a de bonnes raisons de minimiser l'énergie consommée, cela apparaît moins clairement pour ce qui est d'avoir une trajectoire lissée (alors que ce coût donne de bons résultats pour reproduire la dynamique de trajectoires de mains humaines). Il semble enfin qu'il n'y ait pas de modèle qui soit le meilleur dans toutes les situations. L'optimalité semble plutôt à rechercher du côté d'une combinaison de ces différents coûts, ce qui pose le problème de la manière de pondérer chaque critère, d'autant qu'une telle combinaison est appelée à évoluer en fonction des situations.

Dans tous ces exemples effectués en boucle ouverte, la planification qui est faite est indépendante des retours des senseurs, ce qui suppose une dynamique déterministe. Ainsi, les boucles ouvertes sont efficaces pour prédire des trajectoires moyennes mais échouent à prendre en compte le bruit, les délais ou les changements imprévisibles de l'environnement. Au contraire, ceux en boucle fermée montrent de bonnes performances dans ces conditions. Ceux-ci peuvent utiliser un modèle interne permettant d'estimer l'état courant, et les performances sont alors fonction de cet estimateur, car celui-ci permettra d'anticiper les changements avant que les données sensorielles ne soient disponibles. Notons que le parallèle a été effectué avec le cervelet, qui remplit en partie les mêmes fonctions (voir [Wolpert et al., 1998]).

Que ce soit en boucle ouverte ou en boucle fermée, les fonctions de coût sont chaque fois donnée a priori, et supposent donc de pouvoir extraire objectivement les paramètres du comportement moteur qui doivent être optimisés. Il semble cependant exister des situations où les mouvements apparaissent comme suboptimaux au vu des paramètres classiquement optimisés. Dans [Becchio et al., 2008] par exemple, les auteurs observent les trajectoires de sujets devant saisir un objet et le poser sur une zone prédéfinie, ou sur la main d'un humain afin qu'il réalise ensuite un geste avec l'objet. Or le simple fait de donner un objet à un humain, au lieu de le poser sur une table, influence toute la trajectoire y compris la première partie du mouvement destiné à attraper l'objet. La trajectoire semble donc impactée par le contexte social. Cette expérience a été répliquée dans [Lewkowicz et al., 2013] avec des résultats similaires, mais Lewkowicz *et al.* montrent de plus que les cinématiques du mouvement, durant la première partie de la phase d'approche de l'objet, ont des caractéristiques suffisamment différentes pour permettre à un humain de discriminer le contexte (social ou pas). Ainsi, sans voir la seconde partie du mouvement, les expériences menées montrent un taux de succès supérieur au hasard. Notons qu'un simple perceptron (avec une couche cachée) a été capable d'apprendre à différencier les contextes à partir de ces caractéristiques. D'autres études allant dans le même sens ont également été menées, par exemple [Becchio et al., 2010; Ferri et al., 2011].

Ainsi, l'influence de l'intention sociale ou préalable apparaît à très bas niveau dans la réalisation motrice, et une optimisation semble ne pouvoir se limiter aux paramètres moteurs, en dehors de tout contexte psychologique et social.

Enfin, il faut noter que tous les modèles d'optimisation décrits jusqu'ici s'appuient sur une connaissance de la fonction de coût (ou de récompense), et sur une connaissance de la fonction de transition. Or dans le cadre d'une approche développementale, cela ne saurait être le cas. Dans la partie suivante, nous verrons les modèles permettant de palier cela.

### 1.2.3 Apprentissage par renforcement

L'apprentissage par renforcement (RL, [Sutton and Barto, 1998]) concerne les cas dans lesquels ni la fonction de récompense $R$, ni la fonction de transition $T$ ne sont connues à priori. Or, comme nous l'avons vu avec les équations de Bellman, celles-ci sont nécessaires pour déterminer la politique optimale. L'apprentissage par renforcement offre trois manières d'apprendre une politique $\pi$ :

(i) Non basé sur un modèle (model free) : le but est de construire $\pi$ sans construire $T$ ni $R$ ;

(ii) Acteur - critique : cas particulier du cas (i). Le critique représente la fonction de valeur, et l'acteur la politique appliquée. Selon la valeur calculés par le critique, la tendance de l'acteur à effectuer une action va augmenter ou diminuer. La politique est alors améliorée itérativement au cours de l'expérience ;

(iii) Basé sur un modèle (model based) : Le but est alors de construire un modèle de $T$ et de $R$, et à partir de là de calculer la politique optimale (via la programmation dynamique).

Parmi les méthodes non basée sur un modèle, on peut distinguer celle de Monte Carlo ainsi que celles de différence temporelle, qui sont les plus caractéristiques de l'apprentissage par renforcement.

Dans les méthodes de Monte Carlo, le but est d'estimer $V(s)$. La méthode consiste à simuler un grand nombre de trajectoires à partir de chaque état $s$, et à estimer $V(s)$ en moyennant les coûts observés sur ces trajectoires. Afin de pouvoir effectuer ces simulations, les trajectoires doivent être finies. Dans ces méthodes, il faut que l'expérience soit finie pour qu'il y ait un apprentissage : elles ne sont donc pas incrémentales.

Cependant la fonction de valeur peut être mise à jour à la suite de chaque transition, et c'est ce principe qui est à la base des méthodes de différence temporelle (sorte de version "on-line" des méthodes de Monte Cristo). Pour ce faire, au lieu de calculer la fonction de valeur à partir du cumul total espéré, on l'estime à chaque transition en calculant l'erreur par rapport à l'itération précédente.

L'algorithme TD(0) dit de « différence temporelle » compare la récompense obtenue avec celle estimée. Considérons, comme la majeure partie de la littérature sur le sujet, le cas du critère $\gamma$-pondéré suivant : $E[r_0 + \gamma r_1 + \gamma^2 r_2 + ... + \gamma^t r_t + ...|s_0]$. Pour une fonction de valeur basée sur un tel critère, on aurait une mise à jour du type :

$$\textbf{TD(0) :}\ V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)] \tag{1.12}$$

Si cet algorithme permet d'estimer la fonction de valeur (et la preuve de sa convergence a été apportée [Dayan and Sejnowski, 1994]), la fonction de transition reste inconnue. De fait, étant dans un état, l'algorithme ne permet pas de prédire quelle action maximisera la récompense, ce n'est qu'une fois dans l'état qu'il pourra l'estimer.

Afin de remédier à cela, une fonction de valeur $Q$ basée sur les couples $(s, a)$ a été proposée [Watkins, 1989]. $Q^\pi(s, a)$ est la valeur espérée pour le processus partant de s, exécutant l'action a, puis suivant la politique $\pi$. Notons alors que $V^\pi(x) = Q^\pi(x, \pi(x))$. Dans les mêmes conditions, l'algorithme SARSA propose l'équation de mise à jour suivante :

$$\textbf{SARSA : } Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{1.13}$$

Si sa preuve de convergence a également été apportée [Singh et al., 2000], l'inconvénient majeur de cette méthode vient du fait que la mise à jour ne peut se faire que pour l'action réalisée effectivement (du fait de la présence de $a_{t+1}$), et donc l'apprentissage ne peut que suivre la politique courante ("on-policy"), ce qui peut notamment poser des problèmes quant à l'exploration de nouvelles trajectoires.

Afin de permettre une mise à jour en fonction de l'action optimale, indépendamment de l'action réelle effectuée, et donc de la politique suivie ("off-policy"), une mise à jour modifiée a été proposée et constitue l'algorithme de Q-learning [Watkins and Dayan, 1992]. L'équation substitue l'action optimale (présence de $max_a$) à l'action $a_{t+1}$ réellement effectuée, et devient alors :

$$\textbf{Q-learning : } Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{1.14}$$

Enfin, il faut noter que pour pouvoir apprendre autre part qu'à l'état visité par l'agent, et donc accélérer l'apprentissage, il est possible de stocker une mémoire des transitions parcourues et donc de propager l'apprentissage à ces valeurs durant l'apprentissage. Une telle mémoire s'appelle trace d'éligibilité, et a donné lieu aux algorithmes TD($\lambda$), SARSA($\lambda$) et Q($\lambda$). Pour une étude plus approfondie sur la théorie, les applications des processus décisionnels de Markov en intelligence artificielle, et sur l'apprentissage par renforcement, voir les travaux de [Sigaud and Buffet, 2008] dont nous nous sommes inspirés.

Les approches visant à maximiser (ou minimiser) une récompense (ou un coût) sont sujettes aux limitations évoquées dans la partie précédente. Toutefois, celles-ci peuvent être en partie remplacées, ou bien accompagnées d'autres approches complémentaires. Notamment, certaines approches peuvent se concentrer sur la boucle des perceptions et actions propre à l'interaction d'un robot dans son environnement. De part ses propriétés mécaniques, ou encore au vu de la manière qu'aura un robot d'encoder l'information sensorielle, de ses réflexes (issue de l'observation animale ou humaine), ou encore de l'interaction sociale, de nombreuses propriétés peuvent émerger d'apprentissages prenant en compte ces nombreuses dimensions, et peuvent parfois donner lieu à des comportements qu'un observateur extérieur pourrait qualifier de haut niveau. Dans la partie suivante, nous examinerons l'un de ces modèles développés dans le laboratoire où cette thèse a été effectuée, et qui a fortement inspiré notre démarche.

### 1.2.4 Approche PerAc

Le principe clé du modèle Perception-Action (PerAc) [Gaussier and Zrehen, 1995] repose sur l'aspect déterminant de l'interaction entre un robot et son environnement. Plus précisément, il

met en œuvre l'idée que la création d'une représentation sensorielle n'est possible qu'au regard des actions qui en découlent. Il propose un conditionnement réciproque des actions et des perceptions, permettant l'émergence de comportements cohérents avec les besoins du robot et son environnement physique et social.

Le modèle proprement dit est un modèle générique de couplage entre des sensations et des actions permettant de construire des comportements. Inspiré de la counter propagation [Hecht-Nielsen, 1987], PerAc propose le conditionnement d'actions (celui d'une direction à suivre, ou encore d'une configuration motrice par exemple) à partir de l'information sensorielle (la reconnaissance d'un lieu spatial, d'une position particulière d'un membre, d'une expression faciale, *etc*...). La boucle sensorimotrice est refermée par l'environnement, comme le montre la Figure 1.14.

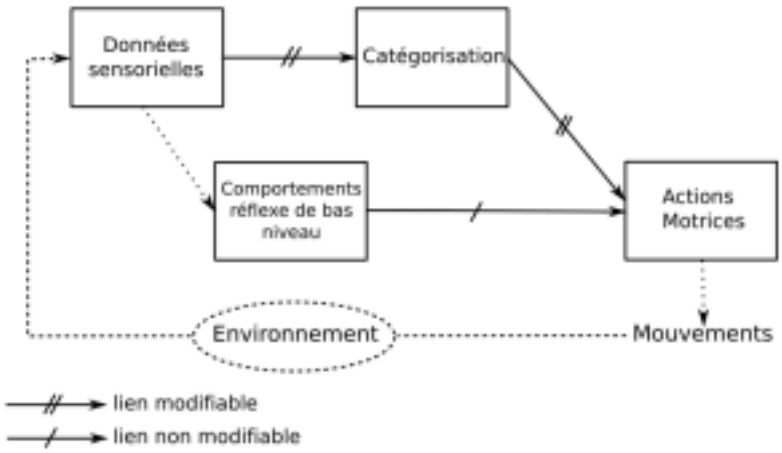


**FIGURE 1.14 –** Boucle sensorimotrice de l'architecture PerAc. Issu de [Boucenna, 2011].

Cette architecture bas niveau permet d'obtenir des comportements sans avoir recours à une représentation symbolique du but. En prenant ses entrées directement dans les capteurs et sans caractérisation symbolique de la situation, le modèle évite le problème de l'ancrage des symboles de l'intelligence artificielle classique ("symbol grounding problem", [Harnad, 1990]).

Une voie réflexe de bas niveau génère des comportements simples (comme ceux liés à la survie) : à partir d'informations sensorielles frustres (infrarouges, tactiles, vision de bas niveau...), des comportements *ad-hoc* comme l'évitement d'obstacles peuvent être générés. Ces comportement peuvent être appris, et ainsi, à partir de perceptions potentiellement de plus haut niveau, des comportements peuvent être anticipés. Cette dynamique permet l'émergence de comportements plus complexes comme la perception d'objet [Maillard et al., 2005] ou le retour vers une source [Giovannangeli and Gaussier, 2008].

Dans le cadre de l'apprentissage d'une tâche de navigation supervisée, l'humain peut donner la direction à suivre comme signal désiré. Le robot apprendra alors à associer l'état courant à la direction désirée. A partir de 3 états (ou plus) dont les actions associées pointent vers un but virtuel, il est possible de construire un bassin d'attraction autour de ce but. La dynamique de reconnaissance des lieux et des actions permet alors au robot de converger naturellement vers le but, sans que celui-ci ne soit explicitement codé [Gaussier and Zrehen, 1995; Gaussier et al., 2000]. Elles peuvent également fournir une stratégie de homing si celles-ci font converger vers

une trajectoire (voir fig. 1.15). L'idée principale ici est que les trajectoires ne sont pas prescrites mais émergent de la dynamique sensorimotrice.

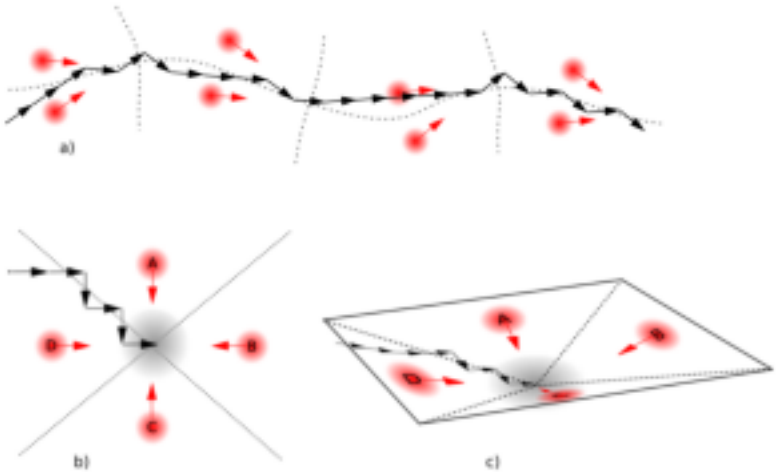

**FIGURE 1.15 –** Chaque lieu est associé à une direction à suivre (flèche rouge). Chaque fois que le robot se trouve dans un lieu, il se tourne automatiquement dans la direction apprise dans ce même lieu. a) représente une trajectoire formée d'associations lieu-action. b) et c) montrent l'apprentissage d'un lieu but avec des associations lieu-action, ainsi que le gradient de convergence que peut suivre un robot pour y arriver. Issu de [Jauffret, 2014].

Dans de grands environnements, ou bien lorsque ceux-ci sont très changeants, de telles stratégies montrent leurs limites. Il n'est notamment pas possible à l'apprentissage de s'adapter pour trouver d'autres chemins, ni de choisir entre plusieurs chemins possibles pour rejoindre un lieu (voir fig. 1.16), sauf à créer des bassins d'attraction différents pour chaque but [Gaussier et al., 2000].

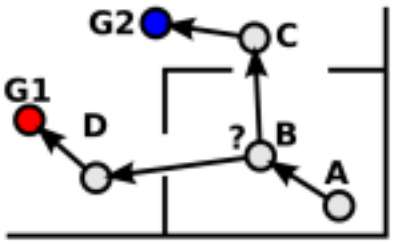


**FIGURE 1.16 –** Selon le but, deux chemins différents peuvent être choisis. Ceci illustre la nécessité d'une planification différente de celle provenant des seules associations lieu-action. Issu de [Hirel, 2011]

Afin de résoudre ce genre de problèmes, une solution consiste à mettre en place des mécanismes de frustration qui détecteront une stagnation dans les progrès du robot à réaliser sa tâche, ce qui permet d'entraîner un changement de stratégie lorsque la frustration est trop grande [Hasson and Gaussier, 2010; Jauffret, 2014]. Une autre solution consiste à utiliser des processus de planification de trajectoires et de sélection d'action, comme ceux basés sur la construction d'une carte cognitive [Banquet et al., 1997].

L'idée de carte cognitive a été émise par Tolman [Tolman, 1948] à partir d'expériences dans lesquelles des rats traversent un labyrinthe. Une carte cognitive correspondrait à une représentation interne de l'environnement permettant à l'animal de trouver des chemins en prenant des raccourcis et en évitant les obstacles. Les cartes cognitives permettent aussi d'expliquer par un apprentissage latent la vitesse à laquelle le comportement du rat s'adapte lorsque une récompense (nourriture) commence à être distribuée. En effet, une telle modification du comportement peut difficilement être expliquée par des apprentissages stimulus-réponse classiques.

La découverte de "cellules de lieu" [O'Keefe and Nadel, 1979], c'est-à-dire des cellules répondant à des endroits précis de l'espace, a permis une meilleure compréhension du fonctionnement de la navigation spatiale chez l'animal. Il a de fait été proposé que la carte cognitive soit présente dans l'hippocampe. La discrétisation de l'espace en cellules discrètes permet, par ailleurs, d'appliquer des modèles probabilistes de sélection de l'action, comme des PDM. Une carte cognitive correspondrait alors à une représentation interne de l'environnement permettant à l'agent d'inférer son chemin parmi plusieurs. Les lieux sont reliés entre eux par des connexions synaptiques, et forment ainsi une carte topologique de l'environnement.

Toutefois, la question se pose alors de savoir ce qui doit être appris. Car si la carte permet de sélectionner le prochain lieu à atteindre, le système ne saurait pour autant déduire l'action à effectuer, puisque celle-ci dépend également de la situation initiale. La solution proposée par [Banquet et al., 1997] consiste à apprendre les transitions en lieux. La topologie étant insuffisante à rendre compte de comportements basés sur la motivation, un mécanisme permet, lorsqu'un but est découvert, de renforcer la connexion entre le neurone qui code la motivation pour rejoindre ce but, et le neurone représentant le lieu courant. Par la suite, l'activité de la motivation se propage à travers la carte cognitive en diminuant à chaque connexion synaptique. L'activité d'un neurone, dans ce modèle, est calculée comme la valeur maximale des activités post-synaptiques qu'il reçoit. Plus le lieu est proche du but plus son activité dans la carte cognitive est grande, et on peut alors trouver le chemin optimal vers le but par une simple remontée de gradient. Ce système est similaire à l'algorithme de Bellman-Ford, calculant le chemin le plus court dans un graphe. De cette manière le robot est à même de remettre en cause la planification de son chemin lors de l'apparition d'un raccourci, ou encore sélectionner le chemin optimal vers un but particulier. Ce modèle a été implémenté et utilisé dans des tâches de navigation robotique [Gaussier et al., 2002; Banquet et al., 2005; Cuperlier et al., 2007; Hirel et al., 2011].

La carte cognitive est donc construite et encodée sous la forme d'un graphe dont les nœuds sont les transitions entres les cellules de lieu, et qui permet de planifier des déplacements dans cet espace topologique. Il est construit en ligne, et la granularité avec laquelle l'espace est pavé par les cellules de lieu peut être contrôlée par un degré de vigilance.

Cette architecture a par la suite été adaptée au contrôle d'un bras robotique [Rolland de Rengerve, 2013]. La configuration motrice du robot, qui lui est donnée par les informations proprioceptives, sera l'espace dans lequel des "cellules de lieu" encoderont certaines configurations. Afin de se déplacer dans cet espace, il ne sera plus question d'orientation du robot, généralement dans l'espace 2D, mais du contrôle du bras dans l'espace de ses degrés de libertés (DDL). Aux associations entre la perception visuelle de la main du robot, et la configuration proprioceptive correspondante, il a été ajouté un modèle du muscle pour contrôler le bras d'une configuration proprioceptive à une autre [De Rengervé et al., 2015]. De plus, ces associations visio-motrices bas niveau permettent, par le maintien de leurs homéostat, des comportements d'imitation immédiate ou différée [de Rengervé et al., 2010; Rolland de Rengerve, 2013]. Plus de détails seront

donnés dans la section 2.1. Mais en reliant ces états, et en apprenant leurs transitions, il est par la suite possible d'obtenir une carte cognitive semblable à celles développées pour la navigation et permettant des comportements motivés. Dans [De Rengervé et al., 2011] une telle carte est par exemple utilisée pour une tâche de "tri de canette", dans laquelle une canette sera déposée à un endroit ou à un autre selon sa couleur. Il en résulte donc une architecture, figure 1.17, qui part de l'encodage bas niveau d'attracteurs sensorimoteurs, et qui permet par la suite de créer une carte cognitive correspondant à des comportements dit plus haut niveau, comme la planification et l'action orientée vers un but. Les comportements pourront être biaisés par une interaction avec le robot, par exemple un signal négatif pourra entraîner un changement du but poursuivi par le robot [De Rengervé et al., 2012].

Les théories décrites jusqu'ici ne sont pas dédiées à l'utilisation d'outils. Afin d'appréhender les spécificités de cette problématique, nous allons, dans la section suivante, proposer un aperçu de différentes approches permettant d'obtenir, ou d'exploiter, la capacité à utiliser des outils.

### 1.2.5 Utilisation d'outils en robotique

Selon Beck [Beck, 1980], l'utilisation d'un outil consiste en l'emploi d'un objet externe de l'environnement, pour altérer efficacement la forme, la position ou l'état d'un autre objet, d'un autre organisme, ou encore de l'utilisateur lui-même.

Dans cette définition répandue de ce qu'est un outil, il est intéressant de noter qu'un outil est par définition extérieur au corps de l'utilisateur. De fait, la problématique de l'utilisation d'outil

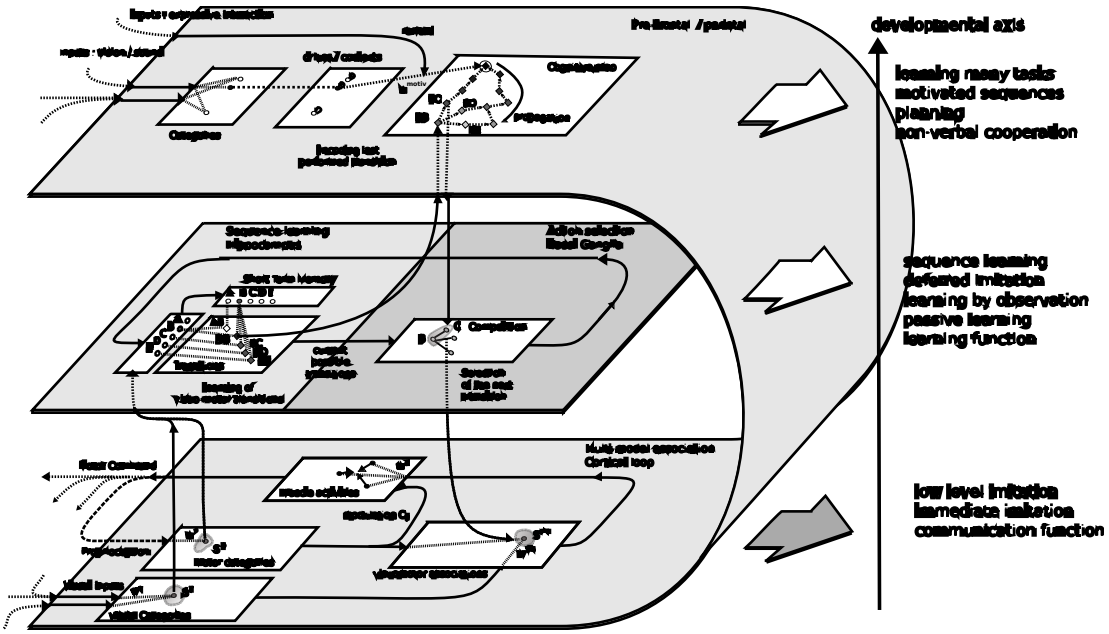


**FIGURE 1.17 –** Modèle issu des travaux de [Rolland de Rengerve, 2013] portant la carte cognitive sur un bras robotique. L'encodage d'associations visuomotrices, que l'on voit en bas de la figure, sert à l'étage supérieur à encoder des transitions entres ces associations, lesquelles, au niveau supérieur, sont utilisées par la carte cognitive qui vient biaiser le choix de ces transitions en fonction de buts.

aura toujours à traiter, à la fois, la question de ce qui constitue le corps, dans sa relation avec la question de l'adaptation à l'environnement.

Du seul point de vue de l'aide que des robots peuvent apporter à des humains, il est à noter que celle-ci serait plus simple à mettre en place si les robots sont capables de s'adapter à un environnement humain, c'est-à-dire fait pour l'homme et non pour la machine. Afin d'assister un humain, un robot devrait donc idéalement être capable de se saisir de la diversité des outils faits pour l'homme, plutôt que d'avoir un outil qui fasse déjà parti de son corps, et donc déjà attaché à son effecteur terminal (voir [Kemp and Edsinger, 2006; Hoffmann et al., 2014]).

Ainsi, comme nous l'avons évoqué plus haut (sec. 1.1.1), l'utilisation d'outil se rapporte à l'adaptation du schéma corporel. Hoffmann et collègues proposent dans [Hoffmann et al., 2010] une revue complète de la question de l'adaptation du schéma corporel lors de l'utilisation d'outils, en robotique.

L'utilisation d'outils soulève également le problème de l'adaptation cinématique. Dans [Rolf et al., 2010], un seul réseau neuronal récurrent est utilisé pour apprendre la cinématique inverse des différents outils, sans réapprentissage ou oubli du schéma corporel. Cependant, la longueur de l'outil est un paramètre de ce réseau neuronal, qui doit être connu à l'avance. Considérant ce problème, dans [Jamone et al., 2012] Jamone et collègues montrent comment la cinématique pourrait être adaptée à l'utilisation de l'outil quand aucune information n'est donnée sur l'outil saisi. Ils ont utilisé pour cela le modèle IMLE qui permet l'apprentissage incrémental de différentes solutions lors d'entrées similaires, et permet ainsi de représenter les relations d'un vecteur d'entrée à un vecteur de sortie dans différents contextes inconnus spécifiques. Pendant l'apprentissage et la phase d'utilisation, l'algorithme traite automatiquement d'un ensemble approprié de modèles linéaires locaux, et cette connaissance est exploitée pour une tâche d'atteinte.

Cependant, Nabeshima et collègues [Nabeshima et al., 2006] défendent l'idée que le véritable problème pourrait être plutôt celui de la détection des altérations corporelles, basées sur les données sensorielles. Ils proposent ainsi un modèle permettant d'adapter le schéma corporel du robot à un outil. Le robot (un bras à 2 degrés de liberté dans le plan, et une caméra) apprend les associations entre l'intégration spatiale et temporelle des informations visuelle et tactile, lorsque la main du robot touche une cible. L'association est ensuite stockée dans une mémoire associative. De cette manière, lorsque le robot touche la cible avec cette fois un outil (un bâton), une nouvelle association entre les informations visuelles et tactiles est apprise dans la mémoire associative, et le contrôleur cinématique du bras est adapté.

La question du schéma corporel peut également être vue comme une étape nécessaire permettant d'accéder à celle d'après, et qui consiste en l'apprentissage des affordances et de leur utilisation. C'est notamment le cas de Fitzpatrick et collègues, qui dans [Fitzpatrick et al., 2003] proposent qu'un robot, après avoir appris à distinguer son corps du reste du monde, apprenne ses affordances à travers les conséquences de ses actions, afin de pouvoir par la suite utiliser ces affordances pour agir en vue d'un but. Notons que ces auteurs insistent sur le fait qu'il est difficile de savoir quand une perception se termine et qu'une action commence, et que la distinction entre les domaines moteur et sensoriel est floue. Ceci se traduit, dans leurs expériences, par le fait que la similarité de deux actions ne soit pas considérée du point de vue de la cinématique, issue des moteurs, mais du point de vue des effets sur le monde extérieur. Cette propriété est alors utilisée pour la compréhension, et l'imitation d'actions effectuées par d'autres.

Ainsi, si l'adaptation du schéma corporel à un outil peut rendre l'utilisateur compétent, dans son usage de cet outil, l'exploitation des affordances peut, elle, fournir un premier pas vers un

usage planifié d'outils. Dans [Stoytchev, 2005], Stoytchev propose qu'un robot apprenne des associations entre, d'une part, un outil et un ensemble prédéfini de comportements, et, d'autre part, un effet, en l'occurrence l'invariant d'un ensemble d'observations. Par la suite, ces associations sont utilisées afin de créer dynamiquement une séquence de comportements permettant de résoudre une tâche.

D'autres travaux reprennent l'idée consistant à utiliser l'apprentissage d'affordances, y compris celles provenant d'outils, afin d'obtenir un but donné, grâce à la connaissance des effets que chaque action pourra, selon le contexte, donner. Notons la similarité de la mécanique à l'œuvre ici avec les schémas, ou schèmes sensorimoteurs de Piaget (voir [Piaget and Cook, 1952]). Dans les deux cas, il s'agit de l'apprentissage d'un contexte dans lequel une certaine action pourra produire un certain résultat. Dans la vision constructiviste de Piaget, l'enfant apprend graduellement, à mesure qu'il interagit avec l'environnement, à coordonner ensemble des schémas bas et hauts niveaux. Comme nous l'avions souligné plus haut (sec. 1.1.1), en soulevant tout autant la question de l'encodage sensorimoteur que celle de la capacité à planifier, la problématique de l'utilisation d'outils soulève un large spectre de questions fondamentales à la robotique développementale (voir [Guerin et al., 2013]).

Une revue des travaux consistant à utiliser les affordances pour l'utilisation d'outils, et permettant une planification, est faite dans [Ugur et al., 2011] (voir aussi les travaux de [Tikhanoff et al., 2013]). Ugur et collègues y classent les différents apprentissages (du type contexte/action/résultat) selon ce qui est appris, et la méthode d'apprentissage utilisée. Dans leurs travaux, ils se positionnent, quant à eux, dans la lignée du principe idéomoteur. En effet, en suivant ce principe, Elsner et Hommel proposent que deux phases soient requises pour l'acquisition de comportements orientés par un but [Elsner and Hommel, 2001]. Dans une première phase, les actions et leurs effets sont associés de manière bi-directionnelle lors d'une phase de babillage moteur. La seconde phase consiste en l'obtention d'un contrôle intentionnel des actions, par la prédiction des effets que ceux-ci peuvent créer.

S'inspirant de ces travaux, Ugur et collègues proposent (dans [Ugur et al., 2011] une approche également en deux phases. Dans la première, une exploration sans but permet à un robot de découvrir les affordances offertes par l'environnement. Elles consistent en l'association bi-directionnelle entre un comportement donné (parmi un répertoire pré-codé de comportements discrets), et un effet, lequel est défini comme la différence entre les caractéristiques (vecteurs de valeurs continues) initiales et finales d'un objet. Dans un second temps, le robot contrôle ses actions par la prédiction de leurs effets. Pour cela, les auteurs tirent parti du fait que les effets et les objets (qui font contexte) soient dans le même espace. Il est de fait possible d'ajouter un effet à l'état en cours, et ainsi, de créer un arbre de recherche dont les noeuds sont composés d'états perceptuels, et les arêtes sont composées de paires comportement-objet. Lorsque le robot a un but donné, il créé un tel arbre, et l'étend graduellement, en partant de l'état qui est à la distance minimale du but. Cela permet au robot de sélectionner une séquence, au sein de cet arbre, lui permettant d'atteindre l'état satisfaisant l'objectif.

Dans [Forestier and Oudeyer, 2016], les auteurs proposent un modèle dans lequel l'espace des senseurs est associé à celui des paramètres permettant le contrôle des moteurs, lequel est effectué à l'aide de DMP (*Dynamical Movement Primitives*, voir [Ijspeert et al., 2013]). Ils proposent un environnement simulé à deux dimensions dans lequel est présent un bras robotique à 4 degrés de libertés, deux outils (des batons de tailles différentes), un objet et des cibles. L'apprentissage consiste à associer les trajectoires motrices, issues des paramètres moteurs, aux résultats

obtenus dans l'espace des senseurs. Ceux-ci sont constitués par les trajectoires de l'effecteur terminal du bras robotique, celles de l'extrémité de chaque outil, et enfin la position finale de l'objet et sa distance à la cible la plus proche. Les auteurs comparent ensuite la progression des résultats lors de tâches consistant à atteindre un but, dans l'espace des senseurs.

Les auteurs testent deux aspects qu'ils estiment être centraux : le premier est la motivation intrinsèque, déterminant l'exploration effectuée, le second est la manière dont est représentée l'encodage sensorimoteur. Pour traiter ce second aspect, ils opposent deux types d'architectures : celle dite plate, et celle dite hiérarchique. Dans l'architecture plate, l'ensemble des senseurs est réuni en un tout, quand dans la seconde architecture cet ensemble est séparé en sous-ensembles, lesquels sont reliés par des liens hiérarchiques. Ils montrent dans leurs travaux que les architecture plates apprennent moins efficacement lorsque l'espace des senseurs a trop de dimensions, comparativement à l'apprentissage hiérarchique, qui est donc privilégié. De la sorte, la question de la motivation intrinsèque se retrouve aussi concernée par le choix du sous-espace, hiérarchiquement situé, de senseurs dont il faut privilégier l'exploration. Par ailleurs, l'importance de cette motivation est d'autant mieux illustrée, dans cet exemple d'utilisation d'outils, que des progrès dans une tâche haut niveau ne sont possibles que si des progrès dans les niveaux inférieurs ont déjà été effectués.

L'apprentissage autonome d'une telle hiérarchisation sensorimotrice n'est pas l'objet d'étude des auteurs, même si ces derniers soulignent l'importance d'un tel apprentissage. Ajoutons que l'on peut se demander si une telle hiérarchisation doit être apprise, en tant que telle, ou bien si celle-ci ne doit pas être émergente, et n'apparaître qu'en tant que solution particulière à un problème particulier. Si l'on part du principe qu'un outil ne devient tel que grâce aux actions qui y sont liées (ses affordances), alors lorsque deux objets sont dans l'environnement, il est, d'un point de vue théorique, tout autant possible que le premier devienne un outil permettant une action sur le second, que l'inverse.

La plupart des résultats décrits jusque maintenant n'ont pas pour objectif premier l'autonomie de dévelopement du robot, ou encore la capacité à résoudre de nouvelles tâches en se basant sur une exploitation nouvelle de précédents apprentissages, même s'il est possible d'exploiter ces méthodes dans un tel objectif. Puisque ces capacités sont liées à notre problématique générale, décrite dans l'introduction, dans la partie suivante nous extrairons plus spécifiquement de la littérature les quelques propriétés générales considérées comme importantes pour obtenir de tels apprentissages.

### 1.2.6 Apprentissage autonome et "open-ended"

La question d'un apprentissage ouvert à différentes finalités, afin de s'adapter aux variations du monde, du robot lui-même et des buts qui l'animent, est une question fondamentale en robotique développementale et autonome. Les nombreux travaux ([Lungarella et al., 2003; Prince et al., 2005; Oudeyer et al., 2007; Stoytchev, 2009; Asada et al., 2009; Law et al., 2014]) portant sur cet apprentissage (dit "open-ended") s'inspirent principalement d'études portant sur le développement de l'enfant, et ont pour objectif de proposer des mécanismes généraux d'apprentissage capables d'intégrer des expériences, de potentiellement très longue durée, au sein d'un ensemble cohérent de compétences, et qui pourront être réutilisées de différentes manières.

D'une part, il faut noter que le développement de l'enfant s'accompagne de changements radicaux et systématiques de son expérience multimodale, et des corrélations sensorielles qui

l'accompagnent. Ceci a mené à l'idée que le développement constitue une extension continuelle des compétences existantes, afin de tenir compte des nouvelles expériences perceptuelles [Smith and Gasser, 2005], ou de l'évolution des dynamiques complexes entre le corps et l'environnement ([Smith and Thelen, 2003]) auxquelles le cerveau de l'enfant doit s'adapter.

En allant plus loin, il a été soutenu que les entrées sensorielles sont initialement non spécifiées, ou "étiquetées", et que le rôle spécifique de chaque modalité doit être inconnu *a priori*. Cette question fondamentale a été illustrée par Denett ([Dennett, 1978]) de la manière suivante : il nous invite à nous imaginer emprisonné dans la salle de contrôle, sans fenêtre, d'un robot géant. Les murs sont recouverts de senseurs, sous la forme de lampes, et d'effecteurs, sans qu'aucun d'entre eux ne soient étiquetés. De plus, peu importe de savoir si derrière le déclenchement d'un effecteur toute une machinerie, potentiellement complexe, se met en branle. Le but final est non seulement de les étiqueter de manière pertinente, mais surtout de contrôler le répertoire de comportements du robot, pour lui permettre par exemple de marcher.

Pierce et Kuipers se sont intéressés à un agent apprenant équipé d'un appareil sensorimoteur "non étiqueté" [Pierce and Kuipers, 1997] : l'agent n'a aucune connaissance *a priori* du système moteur ou des senseurs. L'appareil sensorimoteur se résume alors à des vecteurs de senseurs et de moteurs de données brutes. Les auteurs soutiennent que ces données brutes ne sont pas adaptées pour décrire la structure du monde et prédire les effets de l'action sur les senseurs, et nécessitent donc une étape d'abstraction. Cette abstraction pourra être obtenue en générant des caractéristiques (par exemple en utilisant des ACP), en supposant des relations approximativement linéaires entre les intensités motrices et la dérivé des caractéristiques extraites [Pierce and Kuipers, 1997]. Une hypothèse similaire a été explorée par [Stronger and Stone, 2006], ou encore dans le cadre de la théorie de l'information [Olsson et al., 2006; Lungarella and Sporns, 2006; Kaplan and Hafner, 2005].

D'un autre côté, savoir dans quelle mesure les compétences bas niveaux (comme les associations sensorimotrices) et les capacités dites hauts niveaux (comme la planification de l'action, ou l'utilisation d'outils) partagent des mécanismes d'apprentissage communs demeure une question ouverte. Les dynamiques internes de l'agent et celles de l'environnement, couplées durant l'acquisition de capacités sensorimotrices, pourraient permettre à ces compétences de se développer [Smith and Thelen, 2003], ou encore, de permettre à ces aptitudes bas et hauts niveaux de partager un espace métrique commun [Tani, 2003]. Ainsi il est envisageable que le continuel développement des capacités cognitives et motrices puisse amener à l'émergence de comportements aussi complexes que celui de l'utilisation d'outils [Lockman, 2000]. Et dans une perspective développementale, les comportements bas niveaux pourront former les composantes atomiques qui seront utilisées pour des apprentissages qualifiés de plus hauts niveaux [Stronger and Stone, 2006; Guerin, 2011].

Notons que Rolf et Asada ont quant à eux souligné le rôle crucial des buts [Rolf and Asada, 2015]. Ceux-ci sont souvent manuellement défini par la personne créant le modèle : or ils forment une question centrale pour la question du développement "open-ended", en particulier lorsque l'on considère les cas où ces buts ne pourront être connus en avance, et mettent ainsi en cause l'autonomie de développement d'un robot.

Nous pouvons résumer aux trois points suivants ce que requiert un système qui tendra à être "open-ended" d'un point de vue computationnel :

- Utiliser des fonctions de coût (ou de récompense) qui ne seront pas dédiées à une tâche particulière [Weng et al., 2001; Steels, 2004; Prince et al., 2005].

- Favoriser l'apprentissage de compétences réutilisables, qui pourront être directement applicables dans le cadre de scénarios non expérimentés au préalable (principe de la compositionnalité dans [Cangelosi et al., 2010], voir aussi [Oudeyer et al., 2007; Pezzulo and Castelfranchi, 2009]).
- Acquérir de nouvelles compétences tout en conservant celles acquises aux cours d'expériences passées [Schaal and Atkeson, 1998] (*i.e.*, ce que nous appellerons apprentissage incrémental).

Ces trois points nous semble ainsi devoir faire partie du cahier des charges qu'un modèle, ou qu'un ensemble de modèles interagissants les uns avec les autres, devrait idéalement satisfaire.

## 1.3 Conclusion

Dans la première partie de ce chapitre, nous avons étudié un ensemble de théories psychologiques ou neurobiologiques en parcourant les enjeux soulevés par notre problématique, et plus précisément en examinant différents problèmes que posent l'utilisation d'outils. Nous nous sommes donc intéressé à certaines questions fondamentales touchant au problème du lien entre moteurs et senseurs. Dans la suite de cette première partie, nous avons étendu notre état de l'art aux théories concernant la capacité à faire des séquences. Enfin, nous avons parcouru un ensemble d'autres mécanismes, plus transversaux, desquels nous nous inspirerons dans la suite de ce manuscrit.

Dans la seconde partie de ce chapitre, axée sur les problématique robotiques, nous avons commencé par questionner la question majeure de l'encodage sensorimoteur. Après avoir classé en deux principaux types distincts (qualifiés de "absolu" et "relatif") la manière de faire cet encodage, nous avons brièvement rendu compte des approches incontournables que sont le contrôle optimal, et l'apprentissage par renforcement, et nous avons évoqué certaines limitations qui y sont liées. Nous avons vu que l'approche PerAc permet de poser le problème différemment, en prenant comme facteur premier non un critère à optimiser en vue d'une tâche mais l'apprentissage du lien entre perception et action. Celui-ci, dans sa boucle avec l'environnement, permet un encodage sensorimoteur qui est en même temps une réponse à des contraintes environnementales. Il permet ainsi d'apprendre des réflexes comportementaux, dont l'adéquation avec l'environnement offre des solutions qui peuvent être perçues comme complexes, voire hauts niveaux du point de vue d'un observateur extérieur.

Ainsi cette thèse étudiera différents modèles sensorimoteurs afin d'en examiner les propriétés, et d'en proposer un nouveau à même de répondre à notre problématique, laquelle recouvre aussi bien le problème de l'encodage du schéma corporel, que celui de la capacité à faire des séquences.