Modèles probabilistes formels pour problèmes cognitifs usuels

Pierre Bessière, Julien Diard, Francis Colas

Résumé : Comment un modèle incomplet et incertain de l'environnement peut-il être utilisé pour décider, agir, apprendre, raisonner et percevoir efficacement ? Voici le défi central que les systèmes cognitifs tant naturels qu'artificiels doivent résoudre. La logique, de par sa nature même, faite de certitudes et ne laissant aucune place au doute, est incapable de répondre à cette question. L'approche subjectiviste des probabilités est une extension de la logique conçue pour pallier ce manque. Dans cet article, nous passons en revue un ensemble de problèmes cognitifs usuels et nous montrons comment les formuler et les résoudre avec un formalisme probabiliste unique. Les concepts abordés sont : l'ambigüité, la fusion, la multi-modalité, les conflits, la modularité, les hiérarchies et les boucles. Chacune de ces questions est tout d'abord brièvement présentée à partir d'exemples venant des neurosciences, de la psychophysique ou de la robotique. Ensuite, le concept est formalisé en utilisant un modèle générique bayésien. Enfin, les hypothèses, les points communs et les différences de chacun de ces modèles sont analysés et discutés.

Abstract: How can an incomplete and uncertain model of the environment be used to perceive, infer, decide and act efficiently? This is the challenge that both living and artificial cognitive systems have to face. Symbolic logic is, by its nature, unable to deal with this question. The subjectivist approach to probability is an extension to logic that is designed specifically to face this challenge. In this paper, we review a number of frequently encountered cognitive issues and cast them into a common Bayesian formalism. The concepts we review are ambiguities, fusion, multimodality, conflicts, modularity, hierarchies and loops. First, each of these concepts is introduced briefly using some examples from the neuroscience, psychophysics or robotics literature. Then, the concept is formalized using a template Bayesian model. The assumptions and common features of these models, as well as their major differences, are outlined and discussed.

1 INTRODUCTION

Il est remarquable de constater qu'une large variété de problèmes cognitifs apparaissant fréquemment peut être traitée par un très petit ensemble de modèles fondés sur une approche probabiliste unique. Autrement dit, un petit nombre de constructions mathématiques n'utilisant que le calcul des probabilités peut être appliqué à une grande partie des problèmes rencontrés par les systèmes cognitifs.

Notre objectif dans cet article est de vous en convaincre.[1] Pour cela, nous proposons de passer en revue ces problèmes cognitifs et de décrire pour chacun d'entre eux le modèle bayésien générique permettant de le traiter. Les problèmes cognitifs abordés sont : l'ambigüité, la fusion, la multi-modalité, les conflits, la modularité, les hiérarchies et les boucles. Chacun de ces problèmes est à lui seul une vaste question largement abordée dans la littérature. Qui plus est, chacun est étudié avec des méthodes et des théories divergentes dans différents domaines des sciences cognitives que ce soit, par exemple, en neuroscience, en psychologie expérimentale ou en robotique. Pour chaque domaine et pour chaque problème, des synthèses de littérature existent ou mériteraient d'être faites. Nous n'avons, bien entendu, pas l'ambition de proposer un travail d'une telle ampleur ici. Nous nous limitons à proposer pour chacun des problèmes une courte présentation illustrée par quelques références clefs ou originelles, pour rapidement en venir à la présentation du modèle probabiliste correspondant. Nous espérons ainsi permettre plus de recul, afin de mieux percevoir l'image globale : un emboitement de modèles probabilistes de complexité croissante mais similaires dans leur nature et permettant d'unifier toutes ces questions.

1.1 Les probabilités comme une extension de la logique

Il est frappant de constater qu'un ordinateur a battu le champion du monde d'échec, alors qu'aucun robot n'est capable de faire aussi bien qu'un enfant de cinq ans pour déplacer les pièces sur un échiquier. Comment est-il possible qu'une élite intellectuelle soit systématiquement battue par l'ordinateur quand il s'agit de jouer aux

[1] Cet article est en bonne partie une traduction et une mise à jour d'un article publié en anglais sous le titre *Common bayesian models for common cognitive issues* (Colas et al., 2010).

échecs alors que nous sommes tous capable de ridiculiser ces mêmes machines quand il s'agit de jouer « avec » les échecs ? Quelle différence de nature entre ces deux problèmes peut justifier un tel constat ?

Jouer aux échecs est un problème formel. Tout ce qu'il y a à savoir est parfaitement déterminé : la topologie de l'échiquier, le nombre de pièces, le mode de déplacement et de prise de ces pièces et quelques rares règles complémentaires régissant les débuts et fins de partie. C'est tout. La complexité vient uniquement de la combinatoire.

Jouer « avec » les échecs (percevoir et déplacer les pièces quels que soient leur taille, leur forme, leur aspect, quelle que soit la nature de l'environnement, que ce soit de l'air ou de l'eau, que les pièces soient couvertes de givre ou au contraire brulantes, que la lumière viennent du soleil ou d'un éclairage artificiel, etc.) est un problème d'une tout autre nature. Le problème n'est plus formel, il n'est plus « clos ». La liste des informations nécessaires pour le traiter n'est pas limitée. Certaines sont nécessairement ignorées. Le modèle que nous avons de cette interaction est forcément INCOMPLET. La difficulté vient de cette incomplétude. La majorité d'entre nous est manifestement plus douée que les machines pour surmonter ce deuxième type de difficulté.

Les systèmes cognitifs artificiels comme les êtres vivants doivent résoudre une difficulté fondamentale : COMMENT DECIDER, AGIR, APPRENDRE, RAISONNER ET PERCEVOIR AVEC UN MODELE INCOMPLET ET INCERTAIN DE SON ENVIRONNEMENT ?

Tout modèle d'un phénomène réel est par nature INCOMPLET. Des variables cachées, non prises en compte par le modèle, influencent le phénomène. L'effet de ces variables cachées est que le phénomène et le modèle ne se comportent jamais exactement de la même manière. L'INCERTITUDE est la conséquence directe et inévitable de l'incomplétude. Aucun modèle ne peut prédire exactement les observations futures du phénomène, car ces observations sont « biaisées » par les variables cachées. Pour la même raison, on ne peut pas parfaitement prédire les conséquences d'aucune décision et d'aucune action.

La logique est le paradigme fondateur du raisonnement rationnel. À ce titre, la logique est au cœur de la mathématique, de la démarche scientifique, de la science et de la technologie informatique et est, sans doute, le seul dénominateur commun des théories actuelles de la cognition et de l'esprit. Cependant, de par sa nature même, la logique ne peut pas traiter de l'information incertaine et incomplète. La logique ne peut traiter que de l'information qui est avec certitude, soit vraie soit fausse. Elle ne peut que démontrer des théorèmes, aucune place n'est laissée au doute. L'inférence logique est impossible dès qu'une partie de l'information nécessaire n'est pas disponible.

Les organismes vivants doivent prendre des décisions sensori-motrices avec de l'information manifestement très incomplète. Cependant, les créatures que nous observons survivent de jour en jour en tant qu'individus et ont survécu des millions d'années en tant qu'espèces. Ce simple fait nous démontre qu'il est possible de prendre des décisions sensori-motrices adéquates en dépit de l'incertitude.

1.2 L'approche subjectiviste des probabilités

L'approche SUBJECTIVISTE des probabilités[2] propose les probabilités comme une extension de la logique pour modéliser le raisonnement rationnel avec de l'information incomplète et incertaine (Cox, 1961 ; Jaynes, 2003). Elle s'oppose à l'approche OBJECTIVISTE des probabilités. En simplifiant, les « objectivistes » proposent les probabilités comme un outil pour modéliser le monde (les phénomènes) alors que les « subjectivistes » proposent les probabilités comme un modèle du raisonnement d'un « sujet » à propos du monde. Le raisonnement et les calculs sont tout aussi rigoureux, exacts, inattaquables et « objectifs » dans les deux approches. La « subjectivité » vient uniquement de la référence aux connaissances des sujets. Les subjectivistes reconnaissent et revendiquent que deux sujets peuvent raisonner sur le même phénomène et les mêmes observations avec des CONNAISSANCES PREALABLES différentes. Ils seront alors immanquablement conduits à des conclusions différentes sans qu'aucun des deux n'ait fait d'erreur de raisonnement.

L'approche subjectiviste traite l'incomplétude et l'incertitude avec une démarche en 2 étapes : l'APPRENTISSAGE et l'INFERENCE (voir Figure 1).

[2] Cette approche est souvent appelée de manière partiellement impropre l'approche bayésienne. En effet, il convient d'être prudent avec le terme « bayésien » car différentes communautés ne mettent pas la même définition derrière cet adjectif.

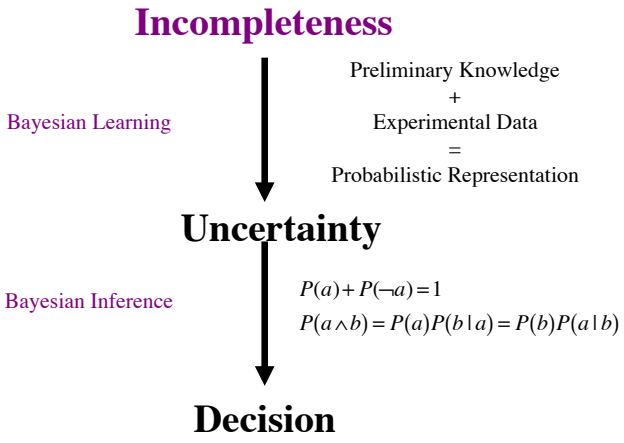

Figure 1 : De l'incomplétude à la décision

L'apprentissage transforme l'irréductible incomplétude des modèles en incertitude quantifiée : des distributions de probabilités. Ces distributions résultent à la fois des connaissances préalables du sujet en train de raisonner et des données expérimentales qu'il recueille en observant les phénomènes étudiés. Les connaissances préalables, mêmes imparfaites et incomplètes, sont pertinentes et donnent un point de vue pour interpréter, organiser et comprendre les données observées. Elles ne sont pas, comme les prémisses en logique, un carcan rigide supposant la complétude. Grâce à certains paramètres laissés libres, elles sont plutôt un canevas prêt à être adapté par les données expérimentales. Les données expérimentales résultent, elles, directement du phénomène observé et reflètent toute la complexité de ce phénomène. Elles dépendent notamment des variables cachées ignorées dans les connaissances préalables. En utilisant les données expérimentales, l'apprentissage prend en compte l'effet de ces variables cachées. Plus important cet effet, plus incertaines, moins piquées, plus grande l'entropie des distributions apprises. C'est ainsi que l'incomplétude est prise en compte et quantifiée.

L'inférence permet des raisonnements complexes avec les distributions de probabilités obtenues par l'apprentissage. Rappelons que les probabilités sont, dans le cas où l'espace sur lesquelles elles portent est discret, un ensemble de nombres positifs dont la somme est l'unité (cette somme devient une intégrale dans le cas d'un espace continu). L'inférence probabiliste ne suppose que deux règles : la règle du produit[3] (Équation 1) et la règle de marginalisation ou règle de la somme (Équation 2).

$$P(X_1,X_2)=P(X_1)P(X_2\mid X_1)=P(X_2)P(X_1\mid X_2) \qquad (1)$$

$$P(X_1)=\sum_{X_2}P(X_1,X_2) \qquad (2)$$

Ces deux règles sont à l'inférence bayésienne ce que le principe de résolution (Robinson, 1979) est à l'inférence logique. Elles sont suffisantes pour n'importe quelle inférence sur des variables discrètes de domaines finis. Ces inférences peuvent être aussi complexes et subtiles que celles effectuées en logique. En fait, il peut être montré que l'inférence logique est un cas particulier de l'inférence probabiliste (Jaynes, 2003).

L'apprentissage lui même peut être vu comme un cas d'inférence probabiliste dans un méta-modèle plus riche considérant les paramètres comme des variables. Les valeurs de ces paramètres sont alors obtenues en maximisant leurs probabilités de façon à expliquer le mieux possible les données expérimentales avec les connaissances préalables (voir Bessière et al (2013), Chapitre 15).

Plus généralement, nous pensons que l'approche subjectiviste des probabilités traite naturellement les problèmes fondamentaux de « signification » et « d'ancrage » que rencontre la logique, grâce, entre autres, à la nature proximale des variables utilisées et à l'apprentissage des modèles probabilistes utilisés. Nous avons discuté en détail de ces questions dans deux articles de cette revue (Bessière et al. 1998a, 1998b).

[3] De cette règle on peut directement dériver le théorème de Bayes : $P(X_2\mid X_1)=P(X_2)P(X_1\mid X_2)/P(X_1)$

1.3 Le formalisme de la programmation bayésienne (*Bayesian Programming*)

L'information nécessaire pour complètement spécifier un modèle probabiliste consiste en : les variables pertinentes, les distributions traduisant les dépendances conditionnelles entre ces variables, les formes paramétriques associées à ces distributions et, finalement les valeurs de ces paramètres.

La manière la plus commune de présenter ce type de modèle dans la littérature est d'utiliser des graphes (voir les RESEAUX BAYESIENS décrits par Pearl (1988) et Jensen (1996) et les MODELES GRAPHIQUES, plus généraux, proposés par Frey (1998) et Jordan (1999)). Les variables pertinentes sont les nœuds, leurs dépendances conditionnelles sont les arcs entre ces nœuds et les formes paramétriques sont associées à chaque nœud.

Comme les probabilités sont une extension de la logique, il est aussi possible d'utiliser un formalisme algébrique pour définir les modèles probabilistes. Nous avons proposé un tel formalisme, appelé « programmation bayésienne » (*Bayesian Programming*) (voir Bessière et al (2003), Lebeltel et al (2004) et Bessière et al (2013) pour une présentation détaillée). Un programme bayésien comprend deux parties :

1. Une DESCRIPTION qui est le modèle probabiliste du phénomène étudié ou du comportement programmé.
2. Une QUESTION qui spécifie les inférences à faire pour résoudre un problème connaissant ce modèle.

La description, elle même, comprend deux parties :

1. Les SPECIFICATIONS qui formalisent les connaissances préalables du modélisateur.
2. L'IDENTIFICATION par laquelle les paramètres laissés libres dans les spécifications sont appris à partir des données expérimentales.

Les spécifications sont constituées de trois parties :

1. Les VARIABLES PERTINENTES retenues pour décrire le phénomène étudié.
2. La DECOMPOSITION qui exprime la distribution conjointe sur les variables pertinentes comme un produit de distributions plus simples (de dimensions réduites). Ces réductions de dimension sont obtenues en exprimant des hypothèses d'indépendances conditionnelles entre variables.
3. Les FORMES PARAMETRIQUES qui associent à chaque distribution apparaissant dans la décomposition soit une forme fonctionnelle soit une question à un autre programme bayésien.

Finalement, une question est définie par la partition de l'ensemble des variables pertinentes en trois sous-ensembles : celui des variables d'intérêts, celui des variables connues et celui des variables libres qu'il va convenir de marginaliser.

Dans cet article nous nous intéresserons essentiellement au choix des variables et à la décomposition retenue pour les modèles envisagés. Nous ne préciserons pas les formes paramétriques qui sont un choix souvent beaucoup plus technique.

Prenons comme exemple la parole[4]. Les variables pertinentes que nous allons retenir sont M (l'ensemble des commande motrices de l'appareil articulatoire), G (les descripteurs de la géométrie de l'articulateur humain) et F (le son produit décrit par ses formants[5]). Nous allons supposer une indépendance conditionnelle entre F et M sachant G. En effet, connaissant la forme du conduit vocal G, nous avons assez d'information pour en déduire le son produit F. Cette hypothèse ne signifie pas, bien évidemment, que les commandes motrices M n'ont pas d'influence sur le son produit G. Elle traduit que cette influence passe par la géométrie du conduit vocal, et que donc si on connait celle-ci, c'est suffisant pour inférer le son. La décomposition ainsi obtenue est alors la suivante (équation 3) :

$$P(M,G,F) = P(M)P(G \mid M)P(F \mid G) \qquad (3)$$

[4] Cet exemple est inspirée de nombreux travaux que nous avons effectué dans ce domaine en utilisant cette approche (Serkhane, 2005 ; Serkhane, et al, 2005 ; Moulin-Frier, 2011 ; Moulin-Frier et al, 2012, 2015 ; Laurent, et al, 2013 ; Laurent, 2014).

[5] Les formants sont les fréquences des pics d'énergie dans l'analyse de Fourier d'un signal acoustique de parole. Les quatre premiers sont largement suffisants pour reconnaitre les sons prononcés.

Nous pouvons alors nous intéresser à un premier problème : quelles commandes motrices sont nécessaires pour émettre un son donné ? Cela se traduit par une question de la forme $P(M \mid F)$. La réponse à cette question peut être calculée de la manière suivante :

$$P(M \mid F) = \frac{P(M,F)}{P(F)} \qquad (4)$$

$$P(M \mid F) = \frac{\sum_G P(M,G,F)}{\sum_F \sum_G P(M,G,F)} \qquad (5)$$

$$P(M \mid F) = \frac{\sum_G P(M)P(G \mid M)P(F \mid G)}{\sum_F \sum_G P(M)P(G \mid M)P(F \mid G)} \quad (6)$$

$$P(M \mid F) = \frac{1}{Z}\sum_G P(M)P(G \mid M)P(F \mid G) \quad (7)$$

où l'équation (4) est obtenue par l'application du théorème de Bayes (équation 1) ; l'équation (5) est obtenue par l'application de la règle de marginalisation (équation 2) à la fois au numérateur et au dénominateur ; l'équation (6) est obtenue en remplaçant la distribution conjointe par sa décomposition (équation 3) ; et où, enfin, l'équation (7) est obtenue en constatant que le dénominateur est une constante de normalisation indépendante de *M* qu'il n'est souvent pas nécessaire de calculer explicitement.

2 AMBIGÜITES ET ILLUSIONS

Les systèmes cognitifs naturels sont plongés dans des environnements riches et variables. Il serait absurde de supposer qu'ils soient capables d'appréhender ces environnements dans toute leur complexité. En conséquence, il ne peut y avoir de réelle bijection entre les états internes du modèle cognitif et les états physiques du phénomène considéré. Autrement dit, un même état interne va correspondre le plus souvent à de multiples états physiques : il est ambigu.

2.1 Problèmes inverses

DESCRIPTION : Un problème est dit INVERSE quand on connait une relation directe sous forme d'une fonction déterministe $y = f(x)$ et que nous recherchons la relation qui à partir de *y* permet d'obtenir *x*. Cette inversion est souvent très difficile, la relation inverse pouvant ne pas avoir de forme analytique et pouvant même ne pas être une fonction dans la mesure où, à un *y* donné, correspondent souvent plusieurs *x* possibles.

EXEMPLES : Les sensations sont l'effet d'un certain phénomène sur les sens. La perception implique d'être capable de récupérer des informations utiles sur le phénomène observé à partir des sensations. La perception peut être vue comme un problème inverse (Poggio, 1984 ; Yuille & Bülthoff, 1996 ; Pizlo, 2001). En effet, on a souvent facilement accès à un modèle direct qui, connaissant le phénomène, décrit les sensations. La perception, est le problème inverse qui, connaissant les sensations, cherche à remonter au phénomène. Par exemple, connaissant la position et la forme d'un objet, on peut calculer son image sur la rétine ; calculer la position et la forme connaissant l'image sur la rétine est toutefois beaucoup plus délicat.

MODELE : Dans la cadre bayésien, un problème inverse est traité en utilisant la symétrie de la règle de Bayes. Dans un exemple de perception générique, appelons Φ une variable décrivant les caractéristiques du phénomène et S les sensations correspondantes. La distribution de probabilité conjointe peut être décomposée ainsi :

$$P(\Phi,S) = P(\Phi)P(S \mid \Phi) \quad (8).$$

Dans cette expression, $P(\Phi)$ est un *a priori* sur le phénomène ; c'est-à-dire ce qu'on sait sur ce phénomène avant la moindre observation. $P(S \mid \Phi)$ est la distribution de probabilité sur les sensations connaissant le phénomène (souvent appelé vraisemblance, quand il est considéré non pas comme une distribution de probabilité sur S mais comme une fonction de Φ) : c'est le modèle direct.

La question probabiliste correspondant à la perception est $P(\Phi \mid S)$. Cette distribution (souvent appelé *a posteriori*) est obtenue par inférence :

$$P(\Phi \mid S) = \frac{P(\Phi)P(S \mid \Phi)}{\sum_{\Phi}\left[P(\Phi)P(S \mid \Phi)\right]} \quad (9)$$

2.2 Problèmes mal posés

DESCRIPTION : Un problème est dit BIEN POSE quand il admet une solution unique. Inversement, un problème MAL POSE admet, soit plusieurs solutions, soit pas de solution du tout. Pour la plupart des problèmes inverses non triviaux, la fonction directe n'est pas injective. La relation inverse n'est donc pas, à proprement parler, une fonction. Quand le modèle direct n'est pas injectif, le modèle inverse est mal posé.

EXEMPLES : La perception est la plupart du temps un problème mal posé. Un exemple bien connu est le cube de Necker (voir Figure 2). La perception de ce cube oscille entre deux solutions concurrentes suivant la face qu'on considère comme étant la plus proche. En fait, ce dessin peut correspondre à la projection en 2D d'une infinité de structures 3D, chacune étant une candidate valide pour la perception.

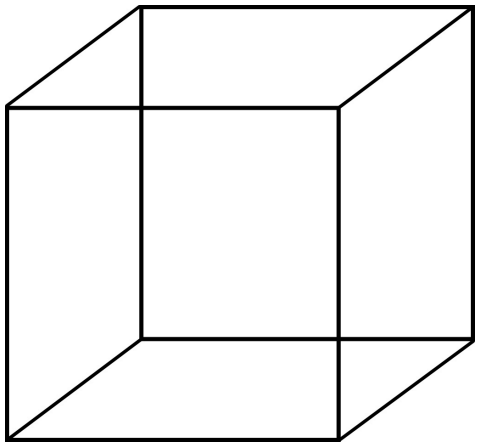

Figure 2 : Le cube de Necker – Un percept bistable

Il y a de très nombreux autres exemples de problèmes de perception mal posés. Retrouver la forme et la position d'un objet connaissant son image sur la rétine est éminemment mal posé. De tels exemples on conduit depuis de nombreuses années à une pléthore de modèles bayésiens de la perception (Knill & Richards, 1996 ; Kersten et al., 2004). Les illusions ont alors une interprétation très simple : elles correspondent au choix de l'une des mauvaises solutions parmi celles possibles (Geilser & Kersten, 2002 ; Laurens & Droulez, 2007 ; Colas et al., 2008).

En robotique, un exemple courant est le *perceptual aliasing*. Quand les capteurs d'un robot lui donnent les mêmes mesures dans deux endroits différents, le problème de la localisation est mal posé et le robot ne peut pas avec certitude, sans autres informations, décider de l'endroit où il se trouve.

Dans le domaine de la parole, associer les sept commandes motrices habituellement considérées pour le conduit vocal (Maeda, 1990) aux quatre premiers formants du signal acoustique est un problème mal posé. Il ne s'agit plus ici d'un problème de perception mais d'un problème de commande puisque ces sept commandes motrices sont celles que l'on doit engendrer pour émettre un son donné. Le fait que ce problème soit mal posé est ici une richesse car il permet de résoudre le problème de la parole même en situation perturbée (par exemple, si on a un

objet dans la bouche). Plus généralement, la commande d'un système moteur redondant (c'est-à-dire, avec plus de degrés de liberté que la dimension de l'espace à accéder) est toujours un problème mal posé.

MODELE : Dans le cadre bayésien, le problème n'est pas de trouver une unique solution satisfaisant toutes les contraintes. Il s'agit d'attribuer une probabilité à chacune des solutions envisageables. L'analogue d'un problème bien posé est une distribution avec un seul mode où une solution est plus probable que toutes les autres. Un problème mal posé correspond au contraire à une distribution avec soit plusieurs modes distincts, soit un « plateau » de solutions équivalentes.

Le problème théorique posé par les problèmes inverses vient de ce que la fonction directe n'admet pas toujours une fonction inverse. Dans le cadre probabiliste, quelque soit la question posée, qu'elle soit directe ou inverse, une solution mathématique peut toujours être trouvée grâce aux deux règles formulées. Certaines questions peuvent, cependant, être très lourdes à calculer. Elles peuvent même demander tant de calcul que le calcul exact et exhaustif est impossible. Il est alors nécessaire de mettre en œuvre des techniques de calcul approché conduisant à une approximation de la distribution de probabilités recherchée. Cette difficulté calculatoire n'est cependant pas directement liée à la nature inverse ou directe de la solution recherchée.

2.3 Discussion

Une ambigüité est la difficulté à trouver une interprétation ou une commande unique à associer à un *stimulus* donné. L'ambigüité est une conséquence directe des problèmes inverses mal posés.

L'ambigüité est souvent résolue en utilisant l'*a priori* $P(\Phi)$. Comme le montre l'équation (8), la vraisemblance est multipliée par l'*a priori* pour calculer l'*a posteriori* recherché. Cela occasionne un tri supplémentaire qui permet éventuellement de choisir une solution parmi plusieurs possibles. L'ambigüité disparait. Malheureusement, elle est éventuellement remplacée par l'illusion. L'illusion apparait quand l'*a priori* élimine les « bonnes » solutions au profit de moins bonnes et conduit à des choix inadéquats.

3 FUSION, MULTI-MODALITE ET CONFLIT

Les systèmes cognitifs naturels utilisent pour percevoir un phénomène donné de très nombreux capteurs. Ils sont plus ou moins redondants suivant leurs proximités de nature et leurs proximités de point de vue. La richesse et la robustesse de la perception viennent de la fusion de ces informations multiples. De nombreuses difficultés peuvent cependant surgir pour effectuer cette fusion. Nous montrons dans ce paragraphe que les concepts de fusion, multi-modalité et conflit peuvent être vus comme différents aspects dans un unique modèle probabiliste. Ce modèle est construit fondamentalement autour d'une hypothèse d'indépendance conditionnelle qui stipule que, connaissant la cause, les observations sont indépendantes. Nous proposons cependant certaines extensions de ce modèle de base qui permettent d'augmenter son expressivité au prix d'une certaine perte de simplicité.

3.1 Fusion

DESCRIPTION : La fusion est le processus qui permet de forger un percept unique à partir de multiples sources d'informations.

EXEMPLES : Un exemple courant de fusion dans l'étude de la perception est appelé la combinaison d'indices (*cue combination*), lorsqu'il s'agit de combiner des informations de la même modalité. Par exemple, Jacobs (1999) étudie la perception de la profondeur d'un cylindre en rotation en utilisant des images de synthèse. Cette profondeur peut être perçue à partir soit du mouvement de la texture, soit de sa déformation, soit de la combinaison de ces deux indices. Weiss et al. (2002) étudient la perception d'un parallélogramme comme la fusion des estimations des vitesses de ses cotés, ce qui peut conduire à diverses illusions comme notamment la perception d'un simple losange. Hillis et al. (2004) proposent un modèle de fusion pour combiner texture et disparité pour percevoir l'orientation d'un plan dans l'espace. Drewing et Ernst (2006) modélisent la perception de la courbure à partir d'informations haptiques en fusionnant des indices de force et de position.

Le modèle de fusion est couramment employé dans les applications technologiques que ce soit, par exemple, en robotique, en diagnostic médical ou pour les filtres qui trient les courriers électroniques.

MODELE : Tous ces exemples utilisent le modèle classique dit de FUSION BAYESIENNE NAÏVE (*naive Bayesian fusion* appelé aussi par certains *weak fusion*). L'hypothèse fondamentale est que chaque information sensorielle est indépendante des autres connaissant le phénomène observé. Autrement dit, c'est l'hypothèse que le phénomène suffit à expliquer toutes les corrélations observées entre les différents capteurs. C'est une hypothèse très forte, bien souvent très fausse, mais la plupart du temps très efficace.

Supposons que $S_1,\cdots,S_N$ soient *N* variables correspondant aux *N* informations capteur à fusionner. Supposons que Φ soit la variable représentant les caractéristiques intéressantes du phénomène. Le modèle de fusion naïve correspond à la décomposition suivante de la distribution conjointe :

$$P\left(\Phi,S_1,\cdots,S_N\right)=P\left(\Phi\right)\prod_{n=1}^{N}\left[P\left(S_n\mid\Phi\right)\right] \quad (10)$$

Il est intéressant de noter que si l'on ne dispose que d'une seule information capteur (*N=1*) ce modèle se ramène au modèle inverse de l'équation (8).

Ce modèle est extrêmement performant pour estimer Φ à partir des informations capteurs connues $s_1,\cdots,s_N$ puisqu'il s'agit essentiellement de faire *N+1* produits :

$$P\left(\Phi\mid s_1,\cdots,s_N\right)=\frac{P\left(\Phi\right)\prod_{n=1}^{N}\left[P\left(s_n\mid\Phi\right)\right]}{P\left(s_1,\cdots,s_N\right)}\propto P\left(\Phi\right)\prod_{n=1}^{N}\left[P\left(s_n\mid\Phi\right)\right] \quad (11)$$

L'hypothèse d'indépendance conditionnelle qui sous-tend ce modèle permet de réduire drastiquement la complexité. La distribution conjointe a une complexité croissant exponentiellement en fonction du nombre de variables, alors que le modèle de fusion naïve a une complexité ne croissant que linéairement en fonction de ce nombre.

La fusion de sources d'informations multiples permet généralement, de plus, un renforcement du signal. Par exemple, dans le cas ou l'a priori $P\left(\Phi\right)$ et les modèles d'observation $P\left(S_n\mid\Phi\right)$ sont des distributions de probabilités gaussiennes, on peut facilement montrer que l'incertitude sur la distribution a posteriori $P\left(\Phi\mid s_1,\cdots,s_N\right)$ est inférieure à toutes celles résultant de chaque capteur séparément.

Dans la littérature, certains des modèles de fusion ne sont pas exactement décrit comme ci-dessus. Ils réfèrent plutôt à une estimation du maximum de vraisemblance (*Maximum Likelihood Estimation* – MLE). Ils cherchent à maximiser la vraisemblance du phénomène définie par $P\left(s_1,\cdots,s_N\mid\Phi\right)$. Pour un jeu de valeurs capteurs connues $s_1,\cdots,s_N$, c'est une fonction de Φ qui quantifie à quel point le phénomène explique les observations. Le MLE sélectionne le phénomène qui explique au mieux les données. La règle de Bayes nous donne :

$$P\left(\Phi\mid s_1,\cdots,s_N\right)=P\left(\Phi\right)P\left(s_1,\cdots,s_N\mid\Phi\right) \quad (12)$$

Il en ressort que l'*a posteriori* sur Φ et la vraisemblance sont proportionnels si et seulement si l'*a priori* sur Φ est uniforme. Le maximum de vraisemblance est donc identique à la fusion si ce n'est qu'il suppose ne pas avoir d'information sur le phénomène considéré en dehors de celle provenant des capteurs.

3.2 Multi-modalité

DESCRIPTION : Les précédents exemples de fusion utilisent une seule modalité sensorielle comme la vision ou le toucher. Cependant, de nombreux modèles portent sur des informations venant de différents sens (ou de différentes technologies de capteurs pour les applications). Ils sont dits alors multimodaux. Les informations venant de sources de natures différentes, elles ne sont pas altérées par les mêmes perturbations et la robustesse globale de la perception du phénomène s'en trouve accrue.

EXEMPLES : La modélisation bayésienne a souvent été utilisée pour étudier la fusion multimodale. Par exemple Anastasio et al. (2000) proposent un modèle multi-sensoriel du renforcement dans le colliculus supérieur. Un renforcement a lieu quand la réponse neurale à un stimulus suivant une modalité est augmentée par un stimulus sur une autre modalité. Dans leur travail, ils montrent de tels renforcements dans le colliculus due à la combinaison de stimuli visuels et auditifs. Zupan et al. (2002) proposent un modèle visuo-vestibulaire de combinaison pondérée d'informations décrite avec le maximum de vraisemblance (MLE). Gepshtein et Banks (2003) étudient l'intégration visuo-haptique en utilisant des indices géométriques de projections pour estimer la fiabilité relative des sources. Kording et Wolpert (2004a) décrivent des expériences d'apprentissage sensori-moteur avec un modèle bayésien d'intégration entre informations visuelles et proprioceptives. Jurgen et Becker (2006) utilisent pour expliquer la perception de rotations du corps la fusion bayésienne d'informations vestibulaire, opto-cinétique, podo-kinesthésique et cognitive. Haith et al. (2008) décrivent une fusion sensori-motrice d'une expérience de pointage visuellement guidée qui prédit un décalage visuel après une adaptation au pointage dans un champ de force.

MODELE : Tous ces modèles de fusion multimodale et de nombreux autres non cités utilisent le modèle de fusion bayésienne naïve décrit plus haut (équations 10 et 11). L'hypothèse est bien toujours que les sensations, qu'elles proviennent de la même modalité ou de différentes origines, sont indépendantes connaissant le phénomène.

Même si le modèle est exactement le même que dans le cas de la fusion monomodale, l'évaluation est souvent beaucoup plus robuste dans le cas multimodal. L'hypothèse d'indépendance conditionnelle n'est jamais totalement valide. Une partie des corrélations éventuelles existantes entre les sensations ne peut pas être expliquée par Φ. Ces corrélations ont cependant plus de chance d'être faibles entre des modalités différentes. En effet, les processus physiques mis en œuvre pour engendrer les sensations à partir du phénomène sont alors de nature complètement différente et ont moins de chance d'être perturbés par une même cause non modélisée. Par exemple, si du brouillard ou un manque de lumière dégradent les différents types d'information visuelle comme la couleur et la profondeur, ils n'affectent pas les informations sonores.

3.3 Conflit

DESCRIPTION : Quand des sources d'information variées sont assemblées, elles peuvent être incohérentes et conduire, chacune, à des percepts éventuellement très différents. Expérimentalement, un conflit est défini comme la situation dans laquelle des comportements différents résultent de la prise en compte des différentes modalités.

Etudier les conflits est très intéressant d'un point de vue expérimental pour mieux comprendre les processus de fusion. Par exemple, l'importance relative des différents indices peut être quantifiée en observant les fréquences d'apparition des comportements en réponse à chacun de ces indices. Dans certains cas, le comportement résultant d'un conflit n'est pas le simple résultat d'un tel choix mais plutôt un autre comportement correspondant précisément à cette situation particulière. Comparer ces comportements aux comportements originaux est alors très riche en enseignements.

EXEMPLES : Pour étudier l'intégration d'informations visuelles et haptiques, Ernst et Banks (2002) ont imaginé des expériences dans lesquelles la hauteur d'une barre peut être estimée, soit à partir de stimuli visuels, soit à partir de stimuli haptiques, soit comme résultat de la fusion des deux. Ces stimuli sont produits par un appareillage qui permet de les rendre incohérents. Leur modèle de fusion bayésienne permet de bien rendre compte des données observées dans les différentes conditions. Battaglia et al. (2003) présentent des expériences avec des conflits entre vision et audition dans des tâches de localisation qu'ils expliquent par un modèle de maximum de vraisemblance. L'effet ventriloque largement étudié par ailleurs (Alais & Burr, 2004 ; Banks, 2004) est un cas particulier de ce type de conflit.

MODELE : Le modèle utilisé pour ce type de conflit est encore une fois celui de la fusion bayésienne naïve (équations 10 et 11). Les conflits apparaissent quand les distributions $P\left(s_n \mid \Phi\right)$ sont notoirement différentes. La distribution $P\left(\Phi \mid s_1, \cdots, s_N\right)$ a alors plusieurs modes bien séparés.

Le concept de conflit est de nature similaire à celui de problème mal posé. Les deux sont définis relativement aux caractéristiques du résultat de l'inférence, pas à celles du modèle. Par contre, l'inversion et la fusion sont des notions définies par la structure du modèle.

3.4 Fusion moins naïve

DESCRIPTION : L'hypothèse de fusion bayésienne considérant les sensations comme indépendantes connaissant le phénomène est parfois manifestement trop forte. Certain modèles rejettent complètement cette hypothèse et ne supposent aucune indépendance conditionnelle (*strong fusion*) ; sauf dans des cas très simples ils conduisent à des modèles trop complexes pour lesquels les inférences sont impossibles dans un temps raisonnable. Divers modèles relâchant partiellement cette hypothèse d'indépendance ont été proposés. Nous présentons ici celui le plus couramment utilisé à base des variables auxiliaires (*ancillary cues*).

EXEMPLES : Landy et al. (1995) proposent ce modèle à variables auxiliaires pour résoudre un problème d'estimation de profondeur. Les variables auxiliaires n'apportent pas directement de l'information sur la profondeur mais permettent de préciser la fiabilité des données sensorielles utilisées dans le processus de fusion. Ils ont appelé ce processus la repondération dynamique (*dynamic reweighting*). Yuille et Bülthoff (1996) proposent la terminologie de couplage fort (*strong coupling*) pour la fusion bayésienne « non naïve ». Ils présentent un exemple de reconstruction de la forme d'un objet à partir d'informations d'ombre et de texture et un exemple de combinaison d'informations de profondeur d'origine binoculaire avec des informations de profondeur monoculaire venant de la parallaxe.

MODELE : Rappelons que le modèle de fusion bayésienne naïve est défini par l'équation (10) :

$$P\left(\Phi,S_1,\cdots,S_N\right)=P\left(\Phi\right)\prod_{n=1}^{N}\left[P\left(S_n\mid\Phi\right)\right]$$

Nous pouvons rajouter une variable auxiliaire A conduisant à la décomposition :

$$P\left(\Phi,A,S_1,\cdots,S_N\right)=P\left(\Phi\right)P\left(A\mid\Phi\right)\prod_{n=1}^{N}\left[P\left(S_n\mid\Phi,A\right)\right] \quad (13)$$

On peut comme précédemment chercher la distribution de probabilités sur le phénomène sachant les sensations, mais elle est obtenue par un calcul différent :

$$P\left(\Phi\mid s_1,\cdots,s_N\right)\propto P\left(\Phi\right)\sum_{A}\left[P\left(A\mid\Phi\right)\prod_{n=1}^{N}\left[P\left(s_n\mid\Phi,A\right)\right]\right] \quad (14)$$

On peut aussi chercher la distribution sur le phénomène sachant à la fois les sensations et la valeur de la variable auxiliaire :

$$P\left(\Phi\mid a,s_1,\cdots,s_N\right)=P\left(\Phi\right)P\left(a\mid\Phi\right)\prod_{n=1}^{N}\left[P\left(s_n\mid\Phi,a\right)\right] \quad (15)$$

Ce modèle est en fait identique au modèle de fusion bayésienne naïve, à condition de considérer que l'on a un phénomène augmenté de la variable auxiliaire $\Phi'=\Phi\wedge A$. La seule différence est alors dans les questions posées. On ne s'intéresse qu'à une partie (Φ) du phénomène (Φ'), soit en connaissant l'autre partie (équation 15) soit en l'ignorant (équation 14).

3.5 Discussion

La fusion est souvent vue comme un produit de modèles probabilistes. La plupart du temps, le résultat d'une inférence sur un modèle de fusion conduit à un résultat proportionnel au produit des résultats qui serait obtenus en faisant l'inférence dans chaque sous-modèle indépendamment :

$$P\left(\Phi\mid s_1,\cdots s_N\right)\propto\prod_{n=1}^{N}P\left(\Phi\mid s_n\right) \quad (16)$$

Toutefois cette proportionnalité n'est pas toujours exacte. Pour qu'elle soit vraie, certaines contraintes doivent être respectées sur la structure et les *a prioris* de chacun des sous-modèles. Si l'on veut garantir cette propriété très intéressante, il faut passer à des modèles de fusion plus complexes utilisant des VARIABLES DE COHERENCE (voir Pradalier et al. (2003) et Bessière et al. (2013, chapitre 8)).

Quand toutes les distributions apparaissant dans le produit sont des distributions gaussiennes, le résultat est aussi une distribution gaussienne. La moyenne de la gaussienne résultat est une somme pondérée des moyennes des gaussiennes du produit. Les poids sont fonction des différentes variances : plus la variance est faible, plus le poids de la moyenne correspondante est fort. De nombreux modèles apparaissant dans la littérature, faisant de telles moyennes pondérées, peuvent être réinterprétés comme des modèles bayésiens de fusion faisant des hypothèses de distributions gaussiennes.

Un conflit découle nécessairement de l'hypothèse qu'il y a un unique phénomène ou objet qu'on est en train d'observer. On espère donc que la distribution $P\left(\Phi \mid s_1,\cdots,s_N\right)$ est monomodale. Quand le conflit est trop important, quand les différentes distributions fusionnées favorisent des valeurs de Φ très différentes, cette distribution devient multimodale. Une manière de résoudre le conflit est alors d'accepter cette multimodalité en remettant en cause l'hypothèse du phénomène unique. Cette situation est souvent qualifiée de SEGMENTATION et on interprète les observations comme des sensations résultant de la perception simultanée de deux phénomènes sous-jacents. Un exemple bien connu est celui de la perception des mouvements de deux objets quand celui au premier plan est transparent et ne masque pas complètement le mouvement de celui en arrière plan $$$ref ?$$$. Pour rendre compte dans un cadre bayésien de la segmentation, des modèles plus complexes doivent être construits. Ces modèles doivent à la fois être capables d'identifier les caractéristiques des différents phénomènes observés et avoir la capacité d'inférer le nombre de phénomènes en jeu. Cette question générique apparait dans la littérature sous des noms divers : *binding*, *pairing*, *unity assumption, causal inference*. Différents modèles bayésiens hiérarchiques ont été proposés en ce sens (voir par exemple Sato et al., 2007).

4 MODULARITE ET HIERARCHIE

Il est difficile de supposer que les systèmes cognitifs naturels traitent leurs entrées sensorielles complexes à l'aide d'une seule couche de traitement. Il est plus probable que ce traitement soit organisé en couches successives qui transmettent des résultats intermédiaires. Cette notion est similaire à la notion de modularité, central à la programmation structurée pratiquée partout pour développer les logiciels de nos ordinateurs. Nous allons dans cette partie présenter successivement différentes manières de structurer les modèles probabilistes : l'appel de sous-programmes, le branchement conditionnel probabiliste, la reconnaissance de modèles et l'abstraction.

4.1 Sous-programme

DESCRIPTION : Le sous-programme est, en programmation classique, l'implémentation la plus évidente de la notion de modularité. Un sous-programme est une unité indépendante, qu'un programme appelle en lui passant des paramètres et dont il récupère les résultats pour continuer son propre calcul. Un sous-programme est donc une ressource pouvant être exploitée pour écrire des programmes plus complexes. De la même façon, il est possible de construire des modèles probabilistes qui puissent servir de ressource pour élaborer des modèles plus complexes.

EXEMPLES : Laskey et Maloney (1997) proposent les fragments de réseaux bayésiens (*network fragments*) à disposition pour être utilisés dans la construction de réseaux plus complexes. Ils utilisent un cadre proche de celui de la programmation objet pour implémenter cette construction. Ils décrivent des applications militaires où un analyste doit modéliser à différents niveaux d'abstraction (unités, pelotons, régiments) pour construire une analyse globale de la situation. Koller et Pfeffer (1997) proposent une approche similaire, les réseaux bayésiens orientés-objet (*object-oriented Bayesian networks*). Les deux souffrent de l'utilisation de modèles graphiques peu propices à la construction de modèles hiérarchiques. Synnaeve et al. (2012 ; 2015) proposent des constructions hiérarchiques beaucoup plus complexes pour gérer les différents niveaux de réflexion nécessaires (déplacement des unités, tactique et stratégique) pour le jeu vidéo Starcraft.

MODELE : Nous proposons dans ce paragraphe de définir l'appel de sous-programme probabiliste en termes algébriques dans le cadre du formalisme de la programmation bayésienne. Supposons que A et B sont des variables pertinentes du modèle global. Ajoutons une nouvelle variable Π qui indexe les différents modèles

envisagés. Attribuons la valeur π_1 au modèle global. Quand nous définissons ce modèle, la valeur de Π est donc connue et fixée à π_1 : la décomposition est donc spécifiée sachant $\left[\Pi = \pi_1\right]$. Par exemple, nous avons :

$$P\left(A,B \mid \left[\Pi = \pi_1\right]\right) = P\left(A \mid \pi_1\right) P\left(B \mid A,\pi_1\right) \quad (17)$$

Un appel à un sous-programme probabiliste peut alors simplement être spécifié en supposant que l'un des termes de la décomposition peut être obtenu en posant une question probabiliste à un autre modèle π_2. Par exemple :

$$P\left(B \mid A,\pi_1\right) = P\left(B \mid A,\pi_2\right) \quad (18)$$

$P\left(B \mid A,\pi_2\right)$, en tant que question probabiliste, peut être le résultat d'une inférence arbitrairement complexe dépendant de la spécification du modèle π_2. De plus, le modèle π_2 peut être « interrogé » par différents autres modèles et peut donc être considéré comme une ressource mise à la disposition des autres modélisateurs, exactement comme un sous-programme en programmation classique.

4.2 Branchement conditionnel probabiliste

DESCRIPTION : Les modèles de mélange sont un moyen de spécifier des distributions multimodales comme une composition de diverses distributions plus simples. La plupart du temps, ces mélanges sont le résultat d'une somme pondérée de distributions monomodales.

EXEMPLES : Les modèles de mélange sont un outil très couramment utilisé pour spécifier des distributions complexes. De très nombreux algorithmes de catégorisation (ou *clustering*) sont disponibles pour apprendre les paramètres (poids) du mélange afin que la distribution résultante décrive au mieux les données observées. Il existe aussi de nombreux algorithmes de classification dont le but est, pour une observation donnée, de retrouver à quelle composante du mélange cette donnée est le plus probablement rattachée. Les exemples les plus classiques de mélange sont les modèles de mixtures de gaussiennes (*Gaussian mixture models, GMM*) qui font une somme pondérée de noyaux gaussiens, mais de nombreux autres noyaux sont aussi utilisés.

Un autre type de modèle de mélange est le « mélange d'expert » (*mixture of experts*). Initialement, cette idée a été introduite dans le cadre des réseaux de neurones (Jacobs et al., 1991). Waterhouse et al. (1996), et beaucoup d'autres à leur suite, ont proposé un algorithme d'apprentissage bayésien des paramètres de tels mélanges. Bishop et Svensen (2003) proposent d'utiliser des méthodes variationnelles pour faire de l'inférence avec des mélanges d'expert et appliquent, par exemple, cet algorithme à la prédiction de la position de l'extrémité d'un bras robotique.

MODELE : Un modèle de mélange est habituellement présenté directement comme une somme pondérée de distributions :

$$P\left(A\right) = \sum_{n=1}^{N} \left[\omega_n P_n\left(A\right)\right] \quad (19)$$

où N est le nombre de composants du mélange, les ω_n sont les poids de chacun des composants $P_n\left(A\right)$ avec la contrainte $\sum_{n=1}^{N} \omega_n = 1$.

Une autre manière de présenter les mélanges consiste à introduire une variable entière Π pouvant prendre N valeurs entre 1 et N. On peut alors construire la décomposition :

$$P\left(A,\Pi\right) = P\left(\Pi\right) P\left(A \mid \Pi\right) \quad (20)$$

où $P(A \mid [\Pi = i]) = P_i(A)$ et où $P([\Pi = i]) = \omega_i$. Le modèle de mélange peut alors être vu comme la réponse à une question probabiliste :

$$P(A) = \sum_{\Pi} [P(\Pi)P(A \mid \Pi)] = \sum_{n=1}^{N} [\omega_n P_n(A)] \quad (21)$$

Cette définition d'un mélange met en évidence le rôle de la variable Π qui est une variable de sélection entre les différentes composantes du mélange. Elle montre aussi clairement le caractère hiérarchique d'une telle construction, dans la mesure où, partant d'un ensemble de modèles simples, on construit un modèle plus complexe grâce à Π.

Un autre avantage important de cette présentation est qu'elle peut être aisément généralisée pour obtenir des branchements conditionnels probabilistes. Supposons que nous ayons une variable C encodant les conditions. On peut alors écrire la décomposition suivante :

$$P(A,\Pi,C) = P(C)P(\Pi \mid C)P(A \mid \Pi) \quad (22)$$

Où $P(C)$ est l'*a priori* sur les conditions, $P(\Pi \mid C)$ est l'expression d'un choix de modèles en fonctions des conditions et $P(A \mid \Pi)$ est la collection de modèles élémentaires utilisables.

La question $P(A \mid C)$ nous permet alors d'exprimer un branchement conditionnel probabiliste :

$$P(A \mid c) \propto \sum_{\Pi} [P(c)P(\Pi \mid c)P(A \mid \Pi)] \quad (23)$$

Si $P(\Pi \mid C)$ est une collection de diracs, alors pour une valeur donnée de *C* un et un seul modèle $P_n(A)$ est retenu et nous avons un sélecteur de modèle très proche d'un branchement non probabiliste. Par contre, si les distributions $P(\Pi \mid C)$ ne sont pas de simples diracs, nous obtenons une généralisation de ce concept de branchement sous la forme d'une somme pondérée des distributions élémentaires dont les poids sont fonction de la valeur *c*.

4.3 Reconnaissance de modèles

DESCRIPTION : Les modèles de mélange permettent la combinaison de différents modèles élémentaires pour en construire un nouveau plus complexe. Le problème dual, qui consiste à choisir parmi plusieurs modèles lequel décrit le mieux les observations, est aussi souvent du plus grand intérêt. Plus généralement, si les différents modèles correspondent à différentes valeurs de paramètres d'un même modèle générique, reconnaitre le modèle revient à trouver les meilleures valeurs de paramètres pour décrire les observations et, *in fine*, se ramène à la question de l'apprentissage machine.

EXEMPLES : Par exemple, Gopnick et Schulz (2004) étudient l'apprentissage des dépendances causales chez les jeunes enfants. Les expériences comportent notamment la question de décider quels objets sont des « blickets » : des objets qui permettent d'allumer une machine lorsqu'on les pose dessus. Les types de réponses fournis par les enfants sont bien prédits par un réseau bayésien causal, même en présence d'*a priori* du type « les blickets sont rares ». La phase d'apprentissage suppose de mettre en compétition plusieurs modèles de dépendance.

Un autre type d'exemples largement développé où la reconnaissance de modèle joue un rôle central sont les modèles de Markov à variables cachées hiérarchiques (*hierarchical* ou *embedded Hidden Markov Models (HMM)*). Ce sont des modèles où un modèle bayésien raisonne pour déterminer, parmi différents HMM, lequel est en train de décrire au mieux la série temporelle d'observation en cours. Nefian et Hayes (1999) ont par exemple proposé un tel modèle hiérarchique pour de la reconnaissance de visage. Chaque sous-modèle est en charge de la reconnaissance d'un trait donné du visage (front, yeux, nez, bouche et menton). Le modèle global assure la reconnaissance du visage en faisant la synthèse des résultats probabilistes des différents sous-modèles. Neal et al. (2003) appliquent les HMM hiérarchiques à la poursuite en 3D de mouvement humain à partir

d'images 2D. La très grande dimensionnalité du problème s'avère plus facile à traiter avec cette approche hiérarchique qu'avec les précédentes méthodes essayées sur cette problématique.

MODELE : Ces différents modèles peuvent être placés dans le cadre général de la reconnaissance bayésienne de modèles. Appelons $\Delta = \left\{\Delta^1, \cdots, \Delta^T\right\}$ la série des observations et Π la variable indexant les modèles en compétition. La décomposition utilisée est :

$$P(\Pi, \Delta) = P(\Pi) P(\Delta \mid \Pi) \quad (24)$$

où $P(\Pi)$ est l'*a priori* sur les modèles et $P(\Delta \mid \Pi)$ est la probabilité d'observation d'une série Δ donnée sachant un modèle, souvent appelée vraisemblance du modèle lorsqu'elle est considérée comme une fonction de Π.

La plupart du temps, on suppose aussi que les observations sont indépendantes et identiquement distribuées (hypothèse i.i.d.) ce qui se traduit formellement par :

$$P(\Delta \mid \Pi) = \prod_{t=1}^{T} \left[P\left(\Delta^t \mid \Pi\right) \right] \quad (25)$$

où les distributions $P\left(\Delta^t \mid \Pi\right)$ sont indépendantes de *t*. Pour chaque valeur possible *n* de Π, la distribution $P\left(\Delta^t \mid \left[\Pi = n\right]\right)$ est un appel au sous-modèle *n* comme défini au paragraphe 4.1.

La question probabiliste correspondant à une reconnaissance de modèle est alors :

$$P(\Pi \mid \delta) \propto P(\Pi) \prod_{t=1}^{T} \left[P\left(\delta^t \mid \Pi\right) \right] \quad (26)$$

Ce type de modèle peut-être qualifié de hiérarchique dans la mesure où l'on utilise un modèle pour raisonner sur un ensemble de sous-modèles.

APPRENTISSAGE DE PARAMETRES : Une extension très courante de ce paradigme de reconnaissance de modèle est l'apprentissage de paramètres. Dans ce cas les différents modèles Π considérés partagent une forme paramétrique commune Π'. Autrement dit, chaque Π est spécifié à l'aide de deux variables Θ et Π'. La décomposition de l'équation (24) devient :

$$P(\Delta, \Pi', \Theta) = P(\Pi') P(\Theta \mid \Pi') P(\Delta \mid \Theta, \Pi') \quad (27)$$

ou, avec l'hypothèse i.i.d. :

$$P(\Delta, \Pi', \Theta) = P(\Pi') P(\Theta \mid \Pi') \prod_{t=1}^{T} \left[P\left(\Delta^t \mid \Theta, \Pi'\right) \right] \quad (28)$$

L'apprentissage des paramètres correspond alors à la question suivante :

$$P(\Theta \mid \delta, \pi') \propto P(\pi') P(\Theta \mid \pi') \prod_{t=1}^{T} \left[P\left(\delta^t \mid \Theta, \pi'\right) \right] \quad (29)$$

Les fonctions de vraisemblances sont complètement définies par les formes paramétriques Π et les paramètres Θ. Cependant, les *a prioris* sur les paramètres $P(\Theta \mid \Pi)$ peuvent éventuellement nécessiter des paramètres supplémentaires appelés hyper-paramètres Ω. La décomposition devient alors :

$$P(\Delta,\Pi",\Theta,\Omega) = P(\Pi")P(\Omega \mid \Pi")P(\Theta \mid \Omega,\Pi")\prod_{t=1}^{T}\left[P\left(\Delta^t \mid \Theta,\Omega,\Pi"\right)\right] \quad (30)$$

PRINCIPES D'ENTROPIE : Tous les principes d'entropie (principe de maximum d'entropie, minimum d'entropie relative, divergence de Kullback-Leibler) et leurs nombreuses applications peuvent être rattachés aux modèles précédents de manière très simple.

Prenons le cas le plus simple défini par l'équation (26) et considérons que nous avons *K* observations possibles (c'est-à-dire, les variables Δ^t peuvent prendre *K* valeurs différentes). En regroupant les observations ayant des valeurs identiques, l'équation (26) devient :

$$P(\Pi \mid \delta) \propto P(\Pi)\prod_{k=1}^{K}\left[P\left(\left[\Delta^t = k\right] \mid \Pi\right)^{n_k}\right] \quad (31)$$

où n_k est le nombre de fois que l'observation *k* a été faite.

En première approximation on peut considérer que $n_k \approx T \times P\left(\left[\Delta^t = k\right] \mid \Pi\right)$ et on obtient :

$$P(\Pi \mid \delta) \propto P(\Pi)\prod_{k=1}^{K}\left[P\left(\left[\Delta^t = k\right] \mid \Pi\right)^{T \times P\left(\left[\Delta^t = k\right] \mid \Pi\right)}\right] \quad (32)$$

Finalement, en supposant un a priori uniforme sur Π et en prenant le logarithme de l'équation (32) on obtient le principe de maximum d'entropie :

$$\log\left(P(\Pi \mid \delta)\right) = T\sum_{k=1}^{K}\left[P\left(\left[\Delta^t = k\right] \mid \Pi\right)\log\left(P\left(\left[\Delta^t = k\right] \mid \Pi\right)\right)\right] + Cst \quad (33)$$

Le modèle le plus probable sachant une suite d'observations est celui qui maximise l'entropie.

4.4 Abstraction

DESCRIPTION : La plupart du temps, le modélisateur utilise l'apprentissage pour choisir un modèle ou un jeu de valeurs de paramètres pour, ensuite, les appliquer au problème à résoudre. En terme probabiliste, l'apprentissage calcule les probabilités des différents modèles (ou jeux de paramètres) puis une DECISION est prise pour n'en retenir qu'un seul (en général le plus probable).

Une autre possibilité est de garder la distribution de probabilité sur les modèles pour l'utiliser dans un méta-modèle. Cette information sur la confiance à attribuer à chaque modèle est en effet précieuse et peut être très utile dans l'ensemble du raisonnement. Cette approche est appelée ABSTRACTION.

EXEMPLES : Diard et Bessière (2008) utilisent l'abstraction pour la localisation de robots. Ils définissent plusieurs CARTES BAYESIENNES correspondant à différents lieux dans l'environnement. Chacune de ces cartes est un modèle probabiliste sensori-moteur de l'interaction du robot avec cette partie de l'environnement. Ils construisent ensuite une carte abstraite basée sur ces cartes élémentaires. Dans cette nouvelle carte, la localisation du robot est obtenue en cherchant quelles cartes bayésiennes correspondent le mieux aux observations sensori-motrices, autrement dit en faisant de la reconnaissance de modèle appliquée aux cartes bayésiennes. Le but étant de naviguer dans l'environnement, une décision sur le lieu exact n'est pas nécessaire. Les ordres moteurs sont déterminés en marginalisant la distribution sur les différentes cartes bayésiennes, profitant ainsi de la riche information sur l'incertitude de la localisation.

Un modèle similaire est utilisé dans le domaine de la perception multimodale sous le nom d'inférence causale (*causal inference*) (Kording et al., 2007 ; Sato et al., 2007). Quand des indices perceptifs sont proches, il est probable qu'ils proviennent de la même source et que les petites différences soient dues à des variables cachées n'ayant pas de rapport avec la position. Par contre, quand ces indices sont très différents, il est beaucoup plus

probable qu'ils proviennent de sources différentes situées en différents endroits. Ne sachant pas le nombre de sources, la stratégie optimale pour déterminer la position de la source ou les positions des sources est alors de garder les différentes hypothèses sur le nombre de sources actives, de marginaliser sur le nombre de sources et de ne décider, *in fine*, que des positions.

MODELE : Soit Π la variable indexant les modèles, soit Δ la variable représentant les observations et soit X la variable d'intérêt. La décomposition considérée est :

$$P(\Pi,\Delta,X) = P(\Pi)P(\Delta \mid \Pi)P(X \mid \Pi) \quad (34)$$

où $P(\Pi)$ et $P(\Delta \mid \Pi)$ sont l'*a priori* et la vraisemblance comme précédemment et où $P(X \mid \Pi)$ explicite la dépendance de la variable d'intérêt en fonction du modèle.

Ce qui est recherché est la distribution sur X sachant les données Δ :

$$P(X \mid \Delta) \propto \sum_{\Pi} \left[P(\Pi)P(\Delta \mid \Pi)P(X \mid \Pi) \right] \quad (35)$$

Appliqué à des classes de modèles et leurs paramètres (c'est-à-dire, quand X est l'ensemble des paramètres Θ) l'abstraction devient ce qui est appelé dans la littérature la sélection bayésienne de modèle (*Bayesian Model Selection*, *BMS*). Kemp et Tenenbaum (2008) utilisent aussi ce même modèle pour la recherche conjointe de la classe de modèle et des paramètres : $P(\Pi,\Theta \mid \Delta)$.

4.5 Discussion

La hiérarchisation est un moyen de contrôler un ensemble de modèles grâce à un méta-modèle en charge d'arbitrer parmi eux. Le méta-modèle utilise les sous-modèles probabilistes comme des ressources de la même façon qu'un programme utilise ses sous-programmes. Deux méthodes sont souvent distinguées dans la littérature : la pondération (*weighing*) et la sélection (*switching*).

La pondération consiste à mélanger les modèles en utilisant des sommes pondérées. Dans le cadre bayésien c'est exactement ce qui est réalisé par le conditionnement probabiliste (voir le paragraphe 4.2). Les poids, éventuellement conditionnés par des variables extérieures, peuvent évoluer dans le temps et occasionner des transitions progressives.

La sélection consiste, en revanche, à choisir parmi les modèles (Stocker & Simoncelli, 2008). Le plus probable est sélectionné et utilisé seul. Cela peut être vu comme une approximation de la somme de l'équation (19) par un seul terme : celui de poids le plus élevé. On mesure bien la perte d'information lors d'une approximation aussi drastique. En contrepartie, les calculs à effectuer sont beaucoup moins couteux. La perte d'information est d'autant plus faible que les poids sont différents. À l'extrême si les poids sont tous nuls sauf un (distribution de Dirac sur les sous-modèles) l'approximation n'en est plus une et les processus de pondération et sélection sont identiques.

Le formalisme bayésien unifie donc ces deux stratégies de hiérarchisation des modèles. La différence entre pondération et sélection, sujet de tant de controverses dans la littérature, ne dépend plus que de la distribution de probabilité utilisée pour « classer » les sous-modèles. Notamment, si cette distribution a une faible entropie, les deux stratégies peuvent être difficilement distinguables.

5 BOUCLES

Il semble très difficile d'admettre que les systèmes cognitifs naturels traitent l'information dans la seule direction allant des entrées sensorielles aux sorties motrices. La neurophysiologie a trouvé d'innombrables exemples de boucles dans le système nerveux de différentes dimensions et de différentes constantes de temps.

5.1 Séries temporelles d'observations

DESCRIPTION : Un système est souvent confronté à une série temporelle d'observations. Ces séries peuvent être utilisées pour divers tâches de calibration, d'estimation, de prédiction, de reconnaissance de séquence, etc. L'estimation de l'état sachant la série d'observation est la question la plus couramment abordée et aussi la plus simple à traiter.

EXEMPLES : Les filtres de Kalman sont les exemples les plus communs de cette catégorie, essentiellement à cause des très fortes contraintes imposées au modèle pour que la solution au problème d'estimation de l'état puisse être résolue analytiquement (Kalman, 1960 ; Ghahramani et al., 1997). Ils sont très largement utilisés en robotique (Thrun et al., 2005), en science de la vie (van der Kooij et al., 1999 ; Kiemel et al., 2002) et dans de très nombreuses applications industrielles. Quand l'espace d'état est supposé discret, on parle alors de modèle de Markov caché (*Hidden Markov Model* ou *HMM*) très commun pour de très nombreuses applications. On se contentera de citer des références « historiques » (Rabiner, 1989 ; Rabiner & Juan, 1993). Quand les hypothèses d'indépendances conditionnelles entre variables ne changent pas dans le temps, on parle de réseaux bayésiens dynamiques (*Dynamic Bayesian Network* ou *DBN)* (Dean & Kanazawa, 1989 ; Murphy, 2002). Ce type de modèles a aussi été très largement traité dans la littérature statistique en utilisant souvent un vocabulaire différent. Plus généralement, toutes les instances de ces modèles de traitement de série temporelles d'observations sont des instances du modèle générique de filtre bayésien (Leonard et al., 1992 ; Bessière et al., 2013).

MODELE : Soit *T+1* variables $O^{0\to T}$ dénotant une série temporelle d'observations de l'instant *0* à l'instant *T*. Soit *T+1* variables d'état $S^{0\to T}$ couvrant la même période. La distribution de probabilité conjointe sur ces variables est décomposée de la façon suivante :

$$P\left(S^{0\to T},O^{0\to T}\right)=P\left(S^0,O^0\right)\prod_{t=1}^{T}\left[P\left(S^t,O^t\mid S^{t-1}\right)\right] \quad (36)$$

Dans cette décomposition est faite une hypothèse de Markov d'ordre 1 : on suppose que l'état à l'instant *t* ne dépend que de l'état à l'instant *t-1*. Autrement dit, on considère que $P\left(S^{t-1}\right)$ est une mémoire suffisante pour le système. Cela ne signifie pas qu'on oublie complètement ce qui est advenu avant, cela veut uniquement dire qu'on considère cette distribution de probabilités comme un résumé suffisant du passé.

L'équation (37), ci-dessous, fait de plus l'hypothèse que la dépendance entre les variables ne change pas au cours du temps (*stationarity hypothesis*) et que cette dépendance contient l'hypothèse que l'observation est indépendante de l'état précédent sachant l'état courant.

$$P\left(S^{0\to T},O^{0\to T}\right)=P\left(S^0,O^0\right)\prod_{t=1}^{T}\left[P\left(S^t\mid S^{t-1}\right)P\left(O^t\mid S^t\right)\right] \quad (37)$$

Dans cette équation $P\left(S^t\mid S^{t-1}\right)$ est souvent appelé modèle de transition ou modèle dynamique. $P\left(O^t\mid S^t\right)$ est, quant à lui, appelé modèle d'observation. Quand ces modèles gardent les mêmes valeurs de paramètres indépendamment du pas de temps *t* considéré, on qualifie le modèle instantané d'homogène ou d'invariant temporellement.

La question la plus courante consiste à chercher à estimer l'état *S* à l'instant *T* sachant toutes les observations jusqu'à cet instant :

$$P\left(S^T\mid o^{0\to T}\right)\propto P\left(o^T\mid S^T\right)\sum_{S^{T-1}}\left[P\left(S^T\mid S^{T-1}\right)P\left(S^{T-1}\mid o^{0\to T-1}\right)\right] \quad (38)$$

On voit que $P\left(S^T\mid o^{0\to T}\right)$ peut être calculé récursivement à partir de $P\left(S^{T-1}\mid o^{0\to T-1}\right)$. On décompose souvent ce calcul en deux phases : une phase dite de prédiction, où on calcule l'état à l'instant *T* en utilisant le modèle dynamique :

$$\sum_{S^{T-1}}\left[P\left(S^T \mid S^{T-1}\right)P\left(S^{T-1} \mid o^{0\to T-1}\right)\right]=P\left(S^T \mid o^{0\to T-1}\right) \quad (39)$$

et une phase d'estimation, où l'on corrige la prédiction précédente en prenant en compte la dernière observation disponible :

$$P\left(S^T \mid o^{o\to T}\right) \propto P\left(o^T \mid S^T\right)P\left(S^T \mid o^{0\to T-1}\right) \quad (40)$$

Le même modèle peut être utilisé pour prédire des états futurs :

$$P\left(S^{T+k} \mid o^{0\to T}\right), k>0 \quad (41)$$

ou pour affiner l'estimation d'un état passé en utilisant les observations obtenues depuis (on parle alors de lissage) :

$$P\left(S^{T-k} \mid o^{0\to T}\right), k>0 \quad (42)$$

La quantité de calcul nécessaire pour résoudre ces deux questions est cependant beaucoup plus importante que pour la question standard de l'équation (38) correspondant à *k=0*. En effet, plus *k* est grand, plus le cout calculatoire est élevé car il est nécessaire de faire des sommes sur *k+1* variables.

On peut aussi utiliser ce modèle pour rechercher la distribution de probabilités sur une série d'états $P\left(S^T, S^{T-1}, \cdots, S^{T-n} \mid o^{0-T}\right)$. C'est typiquement la question qui cherche à résoudre les systèmes de reconnaissance automatique de la parole où les états sont les mots d'une phrase et les observations les phonèmes perçus.

5.2 Copies efférentes

DESCRIPTION : Habituellement dans le cadre de la robotique ou plus généralement du contrôle, on considère non seulement les observations venant de l'environnement mais aussi les actions qui sont appliquées. Dans ce cas les modèles temporels décrits ci-dessus sont enrichis avec des variables d'actions. Ces variables peuvent être « décidées » pour être envoyées comme consignes aux actionneurs mais elles peuvent aussi être « relues » pour voir comment les actionneurs les ont prises en compte. C'est peut-être pour cela qu'on observe des copies efférentes des variables motrices dans les système nerveux des animaux.

EXEMPLES : Les HMM entrée/sortie (*Input/output HMM*) suivent ce type de modèles et sont une référence pour l'inférence grammaticale en traitement du langage (Bengio & Frasconi, 1995). En robotique, la localisation markovienne (*Markov localization*) est la version la plus populaire de ces modèles (Thrun et al., 2005).

En science de la vie, Laurens et Droulez (2007, 2008) appliquent l'estimation bayésienne d'état en utilisant les informations venant des copies efférentes pour modéliser l'estimation de la position de la tête dans l'espace chez l'homme. Ils montrent que ces modèles peuvent rendre compte de nombre d'illusions perceptives constatées chez les pilotes d'essais et les astronautes.

MODELE : Supposons, comme précédemment, que $S^{0\to T}$ soit une série de *T+1* états successifs et que $O^{0\to T}$ soient les variables d'observation correspondantes. Supposons, de plus, que nous avons *T+1* variables de contrôle ou d'action : $A^{0\to T}$. Les variables d'actions sont en général utilisées pour affiner le modèle dynamique :

$$\begin{aligned} &P\left(S^{0\to T}, O^{0\to T}, A^{0\to T}\right) \\ &= P\left(S^0, O^0, A^0\right)\prod_{t=1}^{T}\left[P\left(S^t \mid S^{t-1}, A^{t-1}\right)P\left(O^t \mid S^t\right)P\left(A^t\right)\right] \end{aligned} \quad (43)$$

5.3 Théorie de la décision

DESCRIPTION : Les boucles temporelles présentées jusqu'à maintenant sont utilisées pour estimer les états, qu'ils soient présents, futurs ou passés. Ils peuvent aussi, cependant, être utilisés pour calculer les distributions de probabilités sur les futures actions. Par exemple, on peut chercher l'action ou la série d'action la plus probable pour atteindre un objectif donné. Le but n'est la plupart du temps pas défini en terme probabiliste mais à l'aide de fonctions déterministes qui quantifient quelles actions ou quels états sont favorables ou défavorables. Ces fonctions sont appelées fonctions de cout ou de récompense. Le processus qui consiste à choisir les actions de façon à maximiser les profits ou minimiser les couts définis par ces fonctions est appelé la planification basée sur la théorie de la décision (*decison theoretic planning*).

EXEMPLES : De très nombreux modèles appliquent ce processus à des problèmes de robotique ou de contrôle. Par exemple, quand la localisation markovienne décrite ci-dessus est enrichie avec une fonction de récompense, elle devient un processus de décision markovien partiellement observable (*Partially Observable Markov Decision Process, POMDP*) (voir Kaelbling et al., 1998 ; Boutilier et al., 1999). Si l'état interne S n'est plus « caché » mais peut être mesuré directement, le modèle d'observation $P\left(O^t \mid S^t\right)$ n'est plus utile et est retiré de la décomposition. On a alors à faire à un MDP (*fully observable Markov Decision Process*).

La problématique de la décision (comment passer d'une distribution de probabilités sur une variable à une valeur de cette variable) n'est cependant pas spécifique des boucles. Par exemple, Körding et Wolpert (2004b) cherchent à déterminer quelle fonction de cout est utilisée par les humains dans différentes expériences de contrôle moteur.

MODELE : La fonction de récompense R associe un couple état/action avec un nombre réel qui quantifie son intérêt ou son cout : $R : S^t \times A^t \rightarrow \mathbb{R}$.

La fonction de récompense « pilote » le processus de planification. Le but de la planification est de trouver un plan optimal qui maximise ou minimise une mesure basée sur la fonction R. Cette mesure est le plus souvent l'espérance du cumul de la récompense pondérée :

$$\left\langle \sum_{t=0}^{T} \gamma^t R^t \right\rangle \quad (44)$$

γ^t est un terme d'atténuation qui fait que les récompenses les plus éloignées dans le temps comptent moins que les plus récentes. R^t est la récompense à l'instant t. $\langle \bullet \rangle$ est l'espérance mathématique. Les plans optimisant cette mesure sont appelés « politiques » (*policy*).

Ce processus de planification amène la plupart du temps à des calculs de complexité exponentielle. On en cherche donc des solutions approchées avec des algorithmes d'approximation comme, par exemple, *policy iteration* et *value iteration*.

5.4 Sélection de l'action

DESCRIPTION : Au lieu d'essayer de trouver automatiquement une décomposition structurelle de l'espace d'état, il est possible de réduire la complexité des processus de planification et de contrôle à l'aide d'approches alternatives. On suppose que le modèle contient déjà de l'information à propos de la tâche et du domaine d'application.

Par exemple, modéliser des systèmes cognitifs nécessite d'inclure dans le modèle de l'information sur la structure de l'environnement ou la progression d'une tâche, de manière à créer des filtres élémentaires séparés qui réduisent la complexité et la dimension de l'espace d'état interne. Des systèmes d'attention peuvent aussi être définis pour restreindre les calculs aux seuls sous-ensembles de l'espace d'état jugés pertinents. Enfin des tâches motrices élémentaires peuvent être identifiées et modélisées sous forme de comportements de manière à isoler une partie du cout calculatoire du contrôle du reste de la chaine de raisonnement.

EXEMPLE : Koike (2005) et Koike et al. (2008) ont appliqué cette approche dans le domaine des systèmes sensorimoteurs mobiles autonomes. Koike a montré comment on peut s'accommoder de limites de temps et de mémoire sur des systèmes embarqués grâce à des hypothèses telles que la stationnarité des modèles temporels, l'indépendance partielle entre les processus sensoriels, la sélection du domaine d'intérêt et la sélection de comportement.

MODELES : Incorporer de telles informations dans les modèles les rend moins génériques et plus sophistiqués. L'équation (45) montre la distribution conjointe du modèle complet de Koike (2005).

$$P\left(A^{0\to T},S^{0\to T},O^{0\to T},B^{0\to T},C^{0\to T},\lambda^{0\to T},\beta^{0\to T},\alpha^{0\to T}\mid\pi\right)$$
$$=\left[\prod_{t=1}^{T}\left[\begin{array}{l}\prod_{i=1}^{N_i}\left[P\left(S_i^t\mid S_i^{t-1},A_i^{t-1},\pi_i\right)\right]\prod_{i=1}^{N_i}\left[P\left(O^t\mid S_i^{t-1},C^t,\pi_i\right)\right]\\ P\left(B^t\mid\pi\right)\prod_{i=1}^{N_i}\left[P\left(\beta^t\mid B^t,S_i^t,B^{t-1},\pi_i\right)\right]\\ P\left(C^t\mid\pi\right)\prod_{i=1}^{N_i}\left[P\left(\alpha^t\mid C^t,S_i^t,B^t,\pi_i\right)\right]\\ P\left(A^t\mid\pi\right)\prod_{i=1}^{N_i}\left[P\left(\lambda^t\mid A^t,B^t,S_i^t,A^{t-1},\pi_i\right)\right]\end{array}\right]\right.$$
$$\left.P\left(A^0,S^0,O^0,B^0,C^0,\lambda^0,\beta^0,\alpha^0\mid\pi\right)\right] \quad (45)$$

Dans cette équation π_i fait référence aux N_i filtres élémentaires qui sont supposés indépendants entre eux (justifiant les produits $\prod_{i=1}^{N_i}[\bullet]$). $P\left(S_i^t\mid S_i^{t-1},A_i^{t-1},\pi_i\right)$ sont leurs modèles dynamiques, $P\left(O^t\mid S_i^{t-1},C^t,\pi_i\right)$ leurs modèles capteurs, $P\left(\beta^t\mid B^t,S_i^t,B^{t-1},\pi_i\right)$ leurs modèles de comportement, $P\left(\alpha^t\mid C^t,S_i^t,B^t,\pi_i\right)$ leurs modèles d'attention et $P\left(\lambda^t\mid A^t,B^t,S_i^t,A^{t-1},\pi_i\right)$ leurs modèles de commande motrice. Finalement, $P\left(A^0,S^0,O^0,B^0,C^0,\lambda^0,\beta^0,\alpha^0\mid\pi\right)$ encode l'état initial du système. Ce modèle est alors utilisé pour résoudre une question essentielle pour tout système sensori-moteur : que faire sachant toutes mes observations et actions précédentes ? Formellement, cette question prend la forme suivante :

$$P\left(A^T\mid O^{0\to T},A^{0\to T-1},\lambda^{0\to T},\beta^{0\to T},\alpha^{0\to T},\pi\right) \quad (46)$$

5.5 Discussion

Dans ce paragraphe nous n'avons présenté que des boucles temporelles. Ces boucles sont analogues aux itérations informatiques : le même traitement est exécuté plusieurs fois et le résultat d'un traitement est réutilisé pour le suivant. En biologie les boucles temporelles peuvent avoir des constantes de temps extrêmement différentes, de la nanoseconde à l'échelle moléculaire à plusieurs jours voire années pour certaines régulations ou apprentissages globaux. Ces boucles sont imbriquées les unes dans les autres et interagissent de manière très complexe. Comment réaliser avec les modèles bayésiens de telles imbrications n'est pas encore une question complètement éclaircie.

Les mécanismes itératifs décrits dans ce paragraphe peuvent aussi être étendus à d'autres dimensions que le temps. Il est par exemple possible d'imaginer des modèles spatiaux où un même modèle élémentaire est répliqué plusieurs fois et profite de l'information de ses voisins. Une différence notable avec les filtres temporels naturellement orientés par le temps est que les filtres spatiaux (par exemple, les champs aléatoires de Markov, *Markov Random Fields*) reposent plutôt sur des relations symétriques entre voisins.

6 CONCLUSION

Dans cet article nous avons passé en revue un ensemble de problèmes cognitifs que les systèmes sensori-moteurs tant artificiels que naturels ont à résoudre pour perdurer. Pour chacun de ces problèmes, en utilisant un seul et unique cadre formel probabiliste, nous avons proposé des modèles simples permettant de mieux les comprendre et les résoudre.

Ces problèmes ont été regroupés dans un ordre de complexité croissante, de la fusion et des ambigüités jusqu'au hiérarchies et boucles. Il est à noter que, si les exemples et références viennent au début essentiellement des sciences de la vie, ils viennent pour les plus complexes de la robotique et de l'intelligence artificielle. En effet, les modèles actuellement proposés en biologie sont moins élaborés que ceux utilisés en informatique, même si les choses commencent à changer (Wolpert, 2007 ; Tenenbaum et al., 2011). Mentionnons deux cas emblématiques de cette évolution actuelle. D'une part, dans sa théorie récente, le « principe de l'énergie libre » (*Free Energy Principle*), Karl Friston propose d'interpréter le système cognitif comme une machine probabiliste de traitement de l'information, qui serait guidée par une minimisation de la surprise, c'est-à-dire une maximisation de la probabilité d'avoir prédit les événements observés (Friston et al., 2006 ; Friston & Keibel, 2009). D'autre part, mentionnons les neurosciences probabilistes, qui sont en plein essor et explorent l'hypothèse que les neurones ou assemblées de neurones représentent des distributions de probabilités subjectives (Denève et al., 1999 ; 2001 ; Ma & Jazayeri, 2014 ; Pouget et al., 2013), ce que l'on appelle parfois « l'hypothèse du cerveau bayésien » (*Bayesian Brain Hypothesis*) (Friston, 2010 ; 2012).

Une partie de cette différence entre biologie et informatique vient de ce qu'en biologie un modèle doit être FALSIFIABLE alors qu'en informatique il lui suffit d'être « utile ». L'utilité est beaucoup plus facile et moins rigoureuse à établir que la falsifiabilité. Ceci est plus particulièrement vrai pour les modèles probabilistes. Un modèle probabiliste ne peut pas être à proprement parler falsifiable. Il doit, en fait, nécessairement être comparé à d'autres modèles et la seule « preuve » qui peut être apportée est de démontrer grâce aux outils de comparaisons de modèles décrit au paragraphe 4.3 que l'un de ces modèles décrit mieux les données observées que les autres (c'est-à-dire, il est plus probable sachant les données). Cette méthode de comparaison de modèles reste, pour l'instant, inhabituelle en biologie et est de toute façon délicate à mettre en œuvre.

7 BIBLIOGRAPHIE

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14:257–262.

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–87.

Arulampalam, S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filter for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2).

Banks, M. S. (2004). Neuroscience: what you see and hear is what you get. *Current Biology*, 14(6):236–238.

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, 20(7):1391–1397.

Bengio, Y. and Frasconi, P. (1995). An input/output HMM architecture. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press, Cambridge, MA.

Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E., and Mekhnacha, K. (1998a). Interprétation ou description (i) : Proposition pour une théorie probabiliste des systèmes cognitifs sensori-moteurs. Intellectica, 26:257–311.

Bessière, P., Dedieu, E., Lebeltel, O., Mazer, E., and Mekhnacha, K. (1998b). Interprétation ou description (ii) : Fondements mathématiques de l'approche F+D. Intellectica, 26:313–336.

Bessière, P., Ahuactzin, J.-M., Aycard, O., Bellot, D., Colas, F., Coué, C., Diard, J., Garcia, R., Koike, C., Lebeltel, O., LeHy, R., Malrait, O., Mazer, E., Mekhnacha, K., Pradalier, C., and Spalanzani, A. (2003). Survey: Probabilistic methodology and techniques for artefact conception and development. Technical Report RR-4730, INRIA Rhône-Alpes, Montbonnot, France.

Bessière, P., Laugier, C., and Siegwart, R., editors (2008). *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of STAR. Springer Verlag.

Bessière, P., Mazer, E., Mekhnacha, K. and Ahuactzin, J-M. (2013) *Bayesian Programming*, CRC Press

Bishop, C. M. and Svensen, M. (2003). Bayesian hierarchical mixtures of experts. In Proceedings of the nineteenth conference on uncertainty in artificial intelligence.

Boutilier, C., Dean, T., and Hanks, S. (1999). Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research (JAIR)*, 11:1–94.

Brockwell, P. J. and Davis, R. A. (2000). *Introduction to Time Series and Forecasting* (Second Edition). Springer-Verlag.

Colas, F., Droulez, J., Wexler, M., and Bessière, P. (2008). A unified probabilistic model of the perception of three-dimensional structure from optic flow. *Biological Cybernetics*, pages 132–154.

Colas, F., Diard, J. and Bessière, P. (2010). Common bayesian models for common cognitive issues. *Acta Biotheoretica* 58 :191–216

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150.

Denève, S., Latham, P., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745.

Denève, S., Latham, P., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831.

Diard, J. and Bessière, P. (2008). Bayesian maps: probabilistic and hierarchical models for mobile robot navigation. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of Springer Tracts in Advanced Robotics, pages 153–176. Springer-Verlag.

Drewing, K. and Ernst, M. (2006). Integration of force and position cues for shape perception through active touch. *Brain Research*, 1078:92–100.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–33.

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, 62:1230– 1233.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B*, 364:1211–1221.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology – Paris*, 100:70–87.

Geisler, W. S. and Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuro-science*, 5(6):598–604.

Gepshtein, S. and Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, 13(6):483–488.

Ghahramani, Z., Wolpert, D. M., and Jordan, M. I. (1997). Computational models of sensorimotor integration. In Morasso, P. G. and Sanguineti, V., editors, *Self-organization, computational maps and motor control*, pages 117–47. Elsevier.

Gopnik, A. and Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8(8):371–377.

Haith, A., Jackson, C., Miall, C., and Vijayakumar, S. (2008). Unifying the sensory and motor components of sensorimotor adaptation. In Advances in Neural Information Processing Systems (NIPS 2008).

Harvey, A. C. (1992). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.

Hauskrecht, M., Meuleau, N., Boutilier, L., Kaelbling, L., and Dean, T. (1998). Hierarchical solution of Markov decision processes using macro-actions. In Proceedings of the 14-th Conference on Uncertainty in Artificial Intelligence, pages 220–229.

Hillis, J. M., Watt, S. J., Landy, M. S., and Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *Journal of Vision*, 4:967–992.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, 39:3621–9.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Jaynes, E. T. (2003). *Probability Theory: the Logic of Science*. Cambridge University Press.

Jensen, F. (1996). *An Introduction to Bayesian Networks*. UCL Press.

Jordan, M. I. (1999). *Learning in Graphical Models*. MIT Press. Edited Volume.

Jurgens, R. and Becker, W. (2006). Perception of angular displacement without landmarks: evidence for Bayesian fusion of vestibular, optokinetic, podokinesthetic, and cognitive information. *Experimental Brain Research*, 174:528–543.

Kaelbling, L. P., Littman, M., and Cassandra, A. (1998). Planning and acting in partially observable stochastic domain. *Artificial Intelligence*, 101(1-2):99–134.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, 82(Series D):35–45.

Kemp, C. and Tenenbaum, J. (2008). The discovery of structural form. *Proc. Natl. Acad. Sci. USA*, 105(31):10687–10692.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304.

Kiemel, T., Oie, K., and Jeka, J. (2002). Multisensory fusion and the stochastic structure of postural sway. *Biological Cybernetics*, 87:262–277.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. MIT Press, Cambridge, MA.

Koike, C. (2005). *Bayesian Approach to Action Selection and Attention Focusing. Application in Autonomous Robot Programming*. Thèse de doctorat, Inst. Nat. Polytechnique de Grenoble, Grenoble (FR).

Koike, C., Bessière, P., and Mazer, E. (2008). Bayesian approach to action selection and attention focusing. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of STAR. Springer Verlag.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In Proceedings of the thirteenth conference on uncertainty in artifical intelligence, pages 302–313. Morgan Kaufmann publishers.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, 2(9):e943.

Kording, K. P. and Wolpert, D. M. (2004a). Bayesian integration in sensorimotor learning. *Nature*, 427:244–7.

Kording, K. P. and Wolpert, D. M. (2004b). The loss function of sensorimotor learning. *Proc. Natl. Acad. Sci. USA*, 101(26):9839–9842.

Kuipers, B. J. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, 119(1-2):191–233.

Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, 35:389–412.

Laskey, K. B. and Mahoney, S. M. (1997). Network fragments: representing knowledge for constructing probabilistic models. In Proceedings of the thirteenth conference on uncertainty in artificial intelligence, pages 334–341. Morgan Kaufmann publishers.

Laurens, J. and Droulez, J. (2007). Bayesian processing of vestibular information. *Biological Cybernetics*, 96:389–404.

Laurens, J. and Droulez, J. (2008). Bayesian modeling of visuo-vestibular interactions. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of STAR. Springer Verlag.

Laurent, R., Moulin-Frier, C., Bessière, P., Schwartz, J.-L., and Diard, J. (2013). Integrate, yes, but what and how? A computational approach of perceptuo-motor fusion in speech perception. *Behavioral and Brain Sciences (BBS)*, 36(4):36–37.

Laurent, R. (2014). *COSMO: a Bayesian model of sensorimotor interactions in speech perception*. PhD thesis, Grenoble University.

Lebeltel, O., Bessière, P., Diard, J., and Mazer, E. (2004). Bayesian robot programming. *Advanced Robotics*, 16(1):49–79.

Leonard, J., Durrant-Whyte, H., and Cox, I. (1992). Dynamic map-building for an autonomous mobile robot. *The Intl. J. of Robotics Research*, 11(4):286–298.

Ma, W. J. and Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37:205–20.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Dordrecht: Kluwer.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Moulin-Frier, C. (2011). *Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques : étude, modélisation computationnelle et simulations*. Thèse, Université de Grenoble.

Moulin-Frier, C., Schwartz, J.-L., Diard, J., and Bessière, P. (2011). Emergence of articulatory-acoustic systems from deictic interaction games in a "vocalize to Localize" framework. In *Primate Communication, and Human Language Vocalisation, gestures, imitation and deixis in humans and non-humans*, pages 193–220. John Benjamins.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., and Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor and percep-tuo-motor theories of speech perception: an exploratory Bayesian

modeling study. *Language and Cognitive Processes*, 27,7–8 Special Issue: Speech Recognition in Adverse Conditions:1240–1263.

Moulin-Frier, C., Diard, J., Schwartz, J.-L., and Bessière, P. (2015). Cosmo ("communicating about objects using sensory-motor operations"): a bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley, Berkeley, CA.

Neal, R. M., Beal, M. J., and Roweis, S. T. (2003). Inferring state sequences for nonlinear systems with embedded hidden Markov models. In Thrun, S. and al, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Nefian, A. and Hayes, M. (1999). Face recognition using an embedded HMM. In Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication, pages 19–24.

Pearl, J. (1988). *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pineau, J. and Thrun, S. (2002). High-level robot behaviour control with POMDPs. In AAAI Workshop on Cognitive Robotics.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, 41(24):3141–61.

Poggio, T. (1984). Vision by man and machine. *Scientific American*, 250:106–116.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178.

Pradalier, C., Colas, F., and Bessière, P. (2003). Expressing Bayesian fusion as a product of distributions: Applications in robotics. In Proc. IEEE Int. Conf. on Intelligent Robots and Systems.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 77(2):257–286.

Rabiner, L. R. and Juang, B.-H. (1993). Fundamentals of Speech Recognition, chapter Theory and implementation of Hidden Markov Models, pages 321–389. Prentice Hall, Englewood Cliffs, New Jersey.

Robinson, J. A. (1979). *Logic : Form and Function*. North-Holland, New York, USA.

Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism after effect: Identification of common sources of audiovisual stimuli. *Neural Computation*, 19(12):3335–3355.

Serkhane, J., Schwartz, J.-L., and Bessière, P. (2005). Building a talking baby robot A contribution to the study of speech acquisition and evolution. *Interaction Studies*, 6(2):253–286.

Serkhane, J. E. (2005). *Un bébé androïde vocalisant: Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le développement de la parole*. PhD thesis, Inst. Nat. Polytechnique de Grenoble.

Stocker, A. and Simoncelli, E. (2008). A Bayesian model of conditioned perception. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 1409–1416. MIT Press, Cambridge, MA.

Synnaeve, G. (2012). *Bayesian Programming and Learning for Multi-Player Video Games Bayesian Programming and Learning for Multi-Player Video Games: Application to RTS AI*. PhD thesis, Institut National Polytechnique de Grenoble - INPG.

Synnaeve, G. and Bessière, P. (2015). Multi-scale bayesian modeling for RTS games: an application to starcraft AI. *IEEE Transactions on Computational Intelligence and AI in Games*.

Tenenbaum, J. B. and Kemp, C. and Griffiths, T. L. and Goodman, N. D. (2011) How to grow a mind: statistics, structure, and abstraction. *Science*, 331 :1279–1285

Thrun, S. (2000). Probabilistic algorithms in robotics. *AI Magazine*, 21(4):93–109.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

van der Kooij, H., Jacobs, R., Koopman, B., and Grootenboer, H. (1999). A multisensory integration model of human stance control. *Biological Cybernetics*, 80:299–308.

Waterhouse, S., MacKay, D., and Robinson, T. (1996). Bayesian methods for mixtures of experts. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 351–357. The MIT Press.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Wolpert, D. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, 26:511–24.

Yuille, A. L. and Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In Knill, D. C. and Richards, W.,

editors, *Perception as Bayesian inference*, pages 123–161. MIT Press, Cambridge, MA.

Zupan, L. H., Merfeld, D. M., and Darlot, C. (2002). Using sensory weighting to model the influence of canal, otolith and visual cues on spatial orientation and eye movements. *Biological Cybernetics*, 86(3):209–230.