

Un robot motivé pour apprendre : Le rôle des motivations intrinsèques dans le développement sensorimoteur.

Frederic Kaplan et Pierre-Yves Oudeyer
Sony Computer Science Laboratory Paris
6 rue Amyot 75005 Paris
kaplan@csl.sony.fr, py@csl.sony.fr

Résumé

Cet article présente des travaux récents qui illustrent comment un robot doté d'un système de motivation intrinsèque peut explorer son environnement et apprendre une succession de tâches qui n'ont pas été spécifiées par son programmeur. Un programme générique contrôle le robot et le pousse à rechercher des situations où ses progrès en prédiction sont maximaux. Ces situations, que l'on appelle « niches de progrès », dépendent des opportunités présentes dans l'environnement mais aussi de la morphologie, des contraintes cognitives spécifiques, et de l'expérience passée du robot. Des premiers résultats ont été obtenus dans le domaine de la locomotion, de la découverte des affordances, et des échanges prélinguistiques. Dans chacune de ces expériences, le robot explore les situations « intéressantes » de son point de vue par rapport à ses capacités d'apprentissage et les contraintes de son espace sensorimoteur. L'article discute les résultats de ces premières expériences et conclut sur la possibilité de fournir en retour aux neurosciences et à la psychologie, inspiratrices de ces travaux en robotique, de nouvelles pistes de réflexions et de nouveaux concepts pour penser les processus de développement chez l'enfant.

A robot motivated to learn: Intrinsic motivation and its role in sensorimotor development

Frederic Kaplan and Pierre-Yves Oudeyer

Sony Computer Science Laboratory Paris

6 rue Amyot 75005 Paris

kaplan@csl.sony.fr, py@csl.sony.fr

Résumé

This article presents recent research investigating how a robot equipped with an intrinsic motivation system can explore its environment and learn a sequence of tasks not initially specified by its programmer. A generic software architecture controls the robot, driving it towards situations where learning progress is maximal. These situations – called « progress niches » - depend on opportunities offered by the environment but also on the robot's morphology, specific cognitive biases and past experiences. First results have been obtained in the field of locomotion, affordance discovery and prelinguistic communication. In each of these experiments, the robot explores situations that it evaluated as « interesting » given its learning capabilities and biases of its sensorimotor space. The article discusses the results of these initial experiments and concludes on the relevance of this research to offer neurosciences and psychology – that inspired these investigations in the first place – new kinds of concepts and explanations to think about developmental processes in the young child.

1. Motivation intrinsèque et développement

Le rôle et l'importance des motivations dites « intrinsèques » a été le sujet de plusieurs études depuis les années 1950 en psychologie et en pédagogie (White 1959, Berlyne 1960, Deci 1976, Gibson 1998, Csikszentmihalyi 1991). Les motivations intrinsèques s'opposent aux motivations extrinsèques dans la mesure où leur objet n'est pas la satisfaction de besoins liés à des stimuli extérieurs spécifiques, comme la recherche de nourriture ou de contact social, mais l'attrait de certaines activités pour « elles-mêmes ». Les comportements que l'on attribue d'ordinaire à la curiosité, à la recherche de la nouveauté ou au plaisir de l'exploration seraient le résultat de ce type de motivations. Pour autant, l'articulation précise entre ce type de motivation et le développement du jeune enfant reste largement inconnue. En quelque mois à peine, un enfant apprend à contrôler son corps, à manipuler des objets, à échanger avec ses proches pour devenir un être autonome capable d'interactions complexes tant physiques que sociales. De jour en jour, il acquiert des savoir-faire de plus en plus complexes tant au plan perceptif que dans ses possibilités d'interaction. Quel rôle jouent les motivations intrinsèques dans ce développement continu et ouvert ?

Pour étudier la part des motivations extrinsèques et intrinsèques dans le développement du jeune enfant, il est important de disposer de modèles précis, à partir desquels des expériences peuvent être menées. Le cadre de l'apprentissage par renforcement peut constituer un bon point de départ pour définir plus spécifiquement ce qu'est un système de motivation et son rôle potentiel dans les phénomènes d'apprentissage ouvert (Sutton et Barto 1998). Considérons par exemple un robot capable de percevoir certains aspects de son environnement et devant régulièrement choisir parmi un ensemble de commandes motrices. Dans le cadre de l'apprentissage par renforcement, les robots sélectionnent les actions qu'ils vont effectuer à partir d'un système de valeurs défini par ses programmeurs. Ce système de valeurs définit ce que la machine cherche à obtenir et ce qu'elle cherche à éviter. Il peut par exemple s'agir de rechercher des stimulations tactiles de la part de son propriétaire ou tout simplement des signes de sa présence (son de la voix, détection de

son visage). A l'inverse des chocs trop importants peuvent être associés à des valeurs négatives. Ces principes déterminent tout le comportement de la machine, et sont souvent implémentés sous la forme d'un système de récompenses positives ou négatives qu'elle va tenter d'optimiser. Typiquement, celle-ci va au départ essayer des actions ou des combinaisons d'actions au hasard, et progressivement par « essai-erreur » va sélectionner celles qui maximisent les récompenses positives et minimisent les récompenses négatives. Les actions sélectionnées peuvent être alors aléatoirement modifiées ou combinées pour former de nouvelles stratégies qui sont à leur tour évaluées par le robot. Ainsi, au fur et à mesure de ses expériences, chaque robot peut construire des stratégies propres pour satisfaire ces besoins artificiels définis par son créateur. Dans la plupart des expériences de robotique actuelle, ces situations recherchées sont toujours « extérieures » au robot, et spécifiques d'une tâche particulière. Dans ces conditions, une fois les objectifs atteints (rester près des humains, ne pas de cogner contre les murs) le robot n'a pas de raison de continuer à évoluer. Son développement est précisément limité par le système de motivation extrinsèque dont il a été doté.

C'est pour tenter de palier à cette limite que plusieurs équipes de recherche ont commencé à réfléchir à la manière de doter le robot d'un système de motivation « intrinsèque » tout en restant dans le cadre de l'apprentissage par renforcement (outre nos propres travaux on pourra se référer à Schmidhuber 1991, Huang et Weng 2002, Barto et al 2004). L'idée serait d'équiper le robot d'un système de valeurs qui ne serait pas lié à des tâches spécifiques, définies à l'avance, mais qui le pousserait vers des « situations d'apprentissage ». Ce système le conduirait à explorer les opportunités de son environnement et à découvrir des situations lui permettant de développer de nouvelles compétences. Pour cela il faut formaliser la notion d'apprentissage, de nouveauté ou de progrès d'une manière qui soit directement mesurable par une machine. Des recherches menées en statistiques sur la question de l'« Optimal Experiment Design » (design d'expériences optimales) et en apprentissage artificiel dans le domaine de l'« active learning » (apprentissage actif) se sont penchés sur des questions très proches de celles qui nous intéressent (Fedorov 1972, Thrun 1995). Dans les dix dernières années, plusieurs approches mathématiques ont été proposées pour tenter de définir la meilleure manière d'explorer un espace inconnu de façon à en apprendre le plus possible en un minimum de temps. Ces recherches ont conduit à une amélioration des performances des algorithmes

d'apprentissage. Ces approches constituent une source d'inspiration précieuse pour définir formellement un système de motivation intrinsèque, même si malheureusement elles ne sont pas directement applicables à des robots dotés de capteurs et de senseurs peu précis et qui évoluent dans des espaces réels non-contraints et très inhomogènes.

Cet article présente l'architecture d'un système de motivation intrinsèque adapté aux robots et basé sur le principe de la maximisation du progrès en apprentissage dont le fonctionnement est illustré par un certain nombre d'expériences (voir également Oudeyer et Kaplan 2006, Oudeyer et al. in press). Nous verrons comment ce système générique de motivation permet de rendre compte de trajectoires de développement complexes. Un robot équipé d'un tel système peut apprendre non pas une tâche, mais un ensemble de tâches de complexité croissante.

2. Architecture d'un système de motivation intrinsèque

L'architecture que nous allons décrire est constitué de deux modules (voir figure 1). Le premier module implémente un système de prédiction **M** qui apprend à prédire les conséquences perceptives d'une action entreprise dans un contexte sensoriel et moteur donné. Le second système est un « meta-predicteur » **metaM** qui apprend à prédire les erreurs faites par le prédicteur **M**. En d'autres termes, ce second système modélise le premier et associe à chaque situations reconstruées un niveau de difficulté prédictive.

De façon à doter le robot d'une forme de curiosité, nous avons associé à ces prédicteurs un système de valeurs qui l'incite à éviter les situations trop familières et les situations trop difficiles à prédire, pour privilégier les situations où le progrès en apprentissage est maximal. Il ne s'agit donc pas de choisir les situations pour lesquelles l'erreur en prédiction est minimale, ni celles pour lesquelles elle est maximale, mais celle pour lesquelles l'erreur diminue le plus vite. Il suffit pour cela de calculer la dérivée locale de la courbe d'erreur prédite par **metaM** et correspondant à diverses situations atteignables à partir d'un état sensorimoteur donné, et de choisir une action qui va conduire à la situation dont la dérivée correspondante est la plus négative. Dans ce calcul, il est important de comparer l'erreur obtenue dans des situations similaires. Pour cela, le système de **metaM** utilise un algorithme qui découpe de manière incrémentale l'espace des situations possibles en

groupe de situations similaires. Pour chaque groupe il tient à jour un relevé de l'évolution des erreurs en prédiction, ce qui permet de leur attribuer une valeur (voir figure 2). En outre, de manière à ne pas rester bloquer dans des situations particulières alors que potentiellement d'autres situations intéressantes sont disponibles, des actions aléatoires sont choisies régulièrement. Typiquement, le robot choisit 70 pourcent du temps les situations qu'il juge lui apporter un progrès en apprentissage maximal et 30 pourcent du temps il effectue une action aléatoire. Cela lui permet de mettre à jour son évaluation des différentes situations, et notamment d'en découvrir de nouvelles correspondant à des diminutions maximales de l'erreur en prédiction. Ces situations sont appelées des « niches de progrès ».

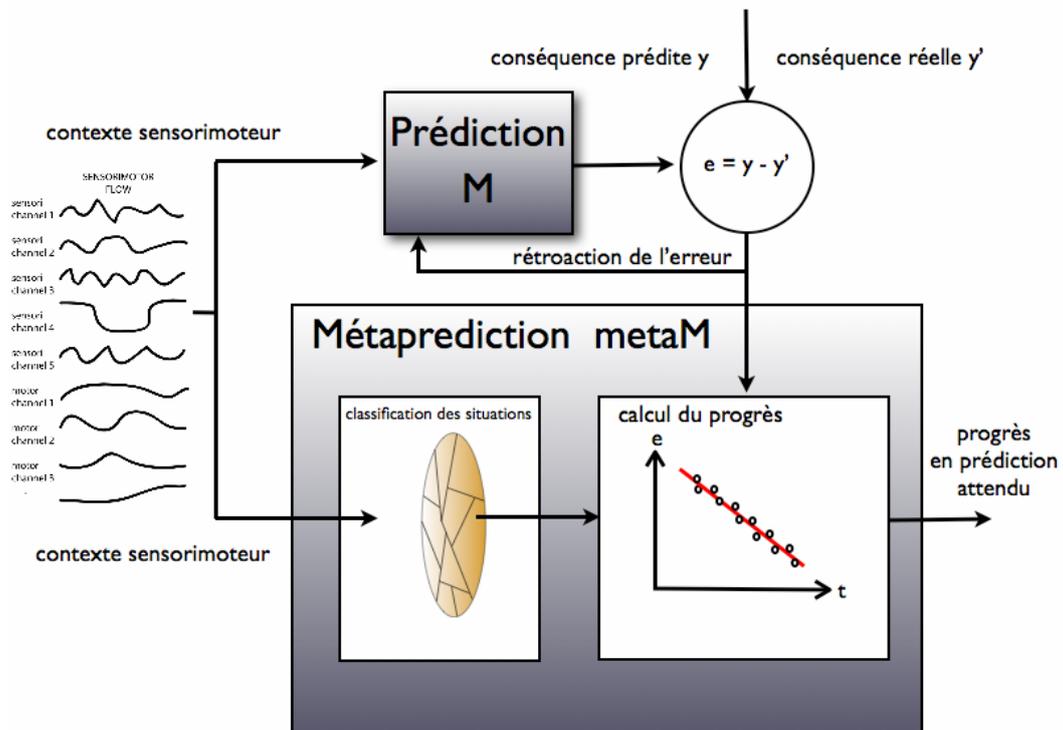


Figure 1 : Le système de motivation intrinsèque est composé de deux modules. : un système de prédiction M et un système de metaprediction metaM.

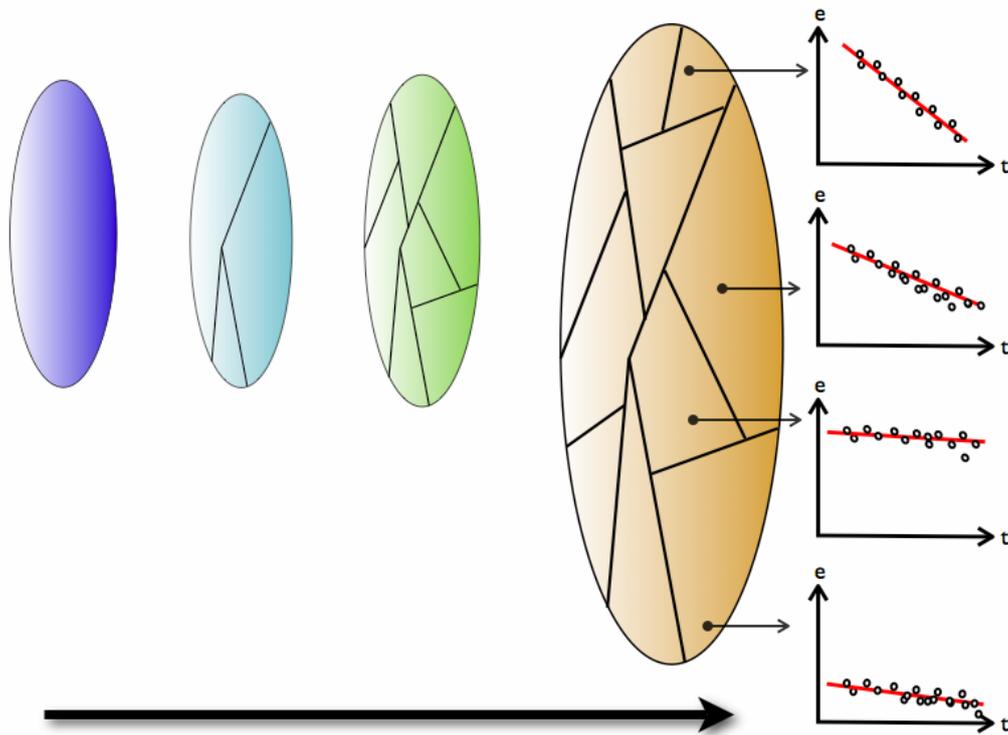


Figure 2: Un algorithme découpe de manière incrémentale l'espace des situations possibles en groupe de situations similaires

Les niches de progrès ne sont pas des propriétés intrinsèques de l'environnement. Elles résultent de la relation entre la structure physique du robot, les biais de ses mécanismes d'apprentissage, ses interactions passées, et l'environnement particulier dans lequel il est placé. Une fois découverte et exploitée, une niche de progrès disparaît au fur et à mesure que la situation à laquelle elle correspond devient plus prédictible. Ainsi, une trajectoire développementale, c'est-à-dire une séquence d'étapes dans lesquelles le robot se focalise sur des activités de complexité croissante, se forme sans qu'elle ait été préprogrammées par le concepteur. Le fonctionnement d'un tel système est illustré sur un exemple abstrait suivant.

Imaginons un environnement dans lequel il existe quatre types d'activités sensorimotrices pour le robot, que l'on appelle contextes sensorimoteurs (e.g. shooter dans une balle, foncer et rebondir dans les murs, courir après le chat, dormir). Si l'on forçait le robot à se concentrer sur chacune de ces activités séparément, on pourrait mesurer l'évolution de son erreur en prédiction dans chacun de ces contextes. Les courbes correspondantes sont

représentées dans la figure 3 (A). On s'aperçoit qu'il y a une situation (1, foncer et rebondir dans les murs) dans laquelle l'erreur reste toujours élevée et ne diminue pas, possiblement parce que cette situation est trop compliquée pour son système d'apprentissage, une autre (4, dormir) dans laquelle elle est toujours basse et ne change pas, et deux situations (2, shooter dans une balle et 3, courir après le chat) pour lesquelles l'erreur est haute au départ, mais diminue ensuite à des rythmes différents. En pratique, le robot est placé dans un environnement où ces différentes activités sont possibles, mais il ne connaît rien des courbes d'apprentissage correspondantes ni même le fait qu'il y a quatre types de contextes spécifiques. Dans ce cas, on doit observer d'abord, une phase d'exploration en grande partie aléatoire qui permet au robot d'une part de s'apercevoir qu'il y a des types de situations différents et d'autre part de se faire une idée initiale de l'intérêt en terme d'apprentissage de chacune de ces situations. On doit alors observer le comportement illustré par le graphique 3 (B). Le robot évite de se trouver dans les situations 1 et 4 car elle ne permettent pas de progrès en apprentissage. Il les explore cependant de temps en temps et par hasard, ce qui lui permet de vérifier qu'elles restent peu intéressantes. A l'inverse, il va se concentrer d'abord sur la situation 3 qui est celle pour laquelle ses prédictions s'améliorent le plus vite initialement. Après une première période, la situation 3 est maîtrisée et prédictible : le robot se met alors spontanément à explorer la situation 2 qui à ce stade de son développement lui procure le plus de progrès en apprentissage.

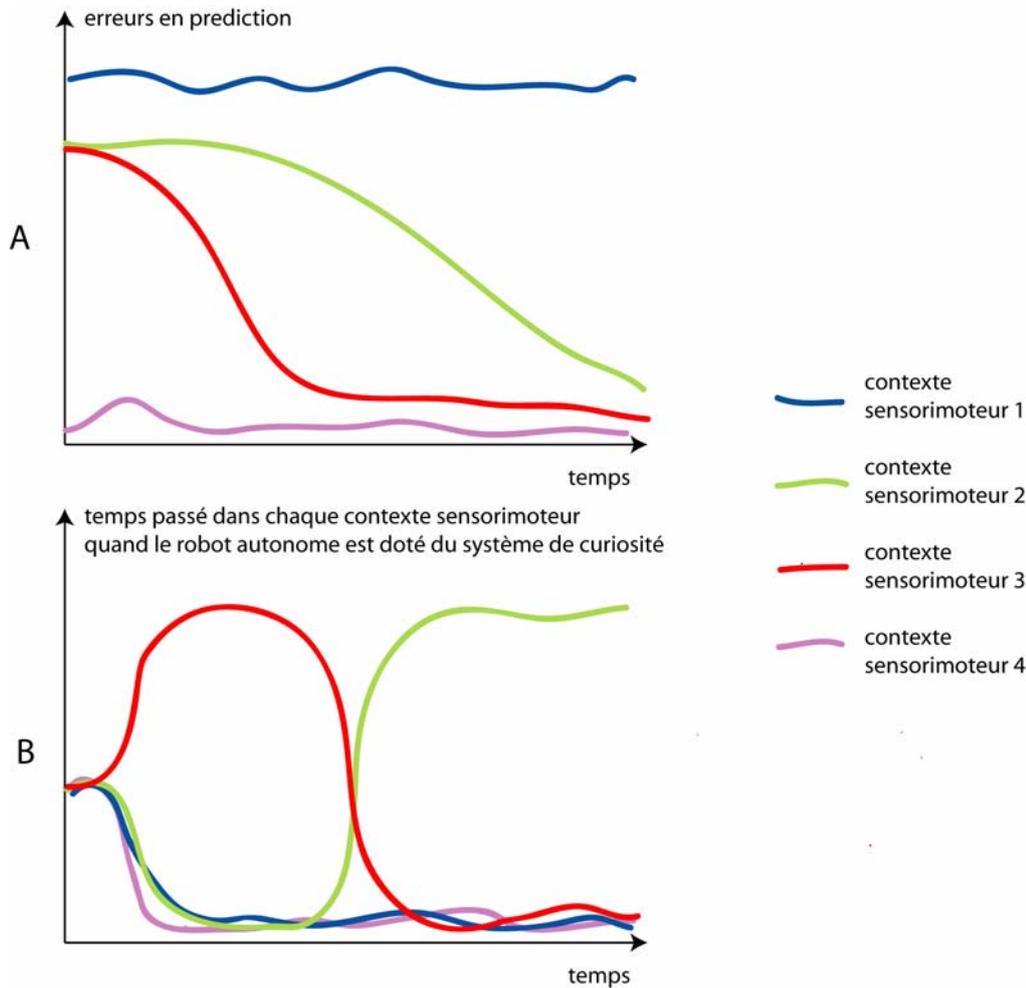


Figure 3 : Exemple abstrait avec quatre types d'activités sensorimotrices plus ou moins facile à apprendre. Le robot évite de se trouver qui ne permettent pas de progrès en apprentissage (trop facile ou trop difficile) pour se concentrer sélectivement sur les situations où son apprentissage progresse de façon optimale.

3. Expérience 1 : Apprentissage de la marche

L'architecture que nous venons de décrire est générique dans la mesure où elle peut être appliquée à n'importe quel espace de senseurs et de moteurs qu'un robot peut explorer, et par définition n'est pas spécifique d'une tâche particulière. Considérons un premier exemple dans lequel un robot AIBO contrôle le mouvement de ses différents moteurs en appliquant des signaux sinusoïdaux (voir figure 4). Pour chaque moteur, il fixe la période, la phase et l'amplitude du signal. Son système de prédiction tente de prédire les effets de ces différents jeux de paramètres sur la manière dont l'image captée par sa caméra est

modifiée, ce qui reflète indirectement le mouvement de son buste (pour lequel il n'a pas de capteur). A chaque itération de l'algorithme, le robot choisit la valeur du prochain jeux de paramètres à essayer de façon à maximiser la réduction de l'erreur en prédiction, sauf dans 30 pourcent des cas où comme on l'a expliqué plus haut il choisit des paramètres aléatoires.

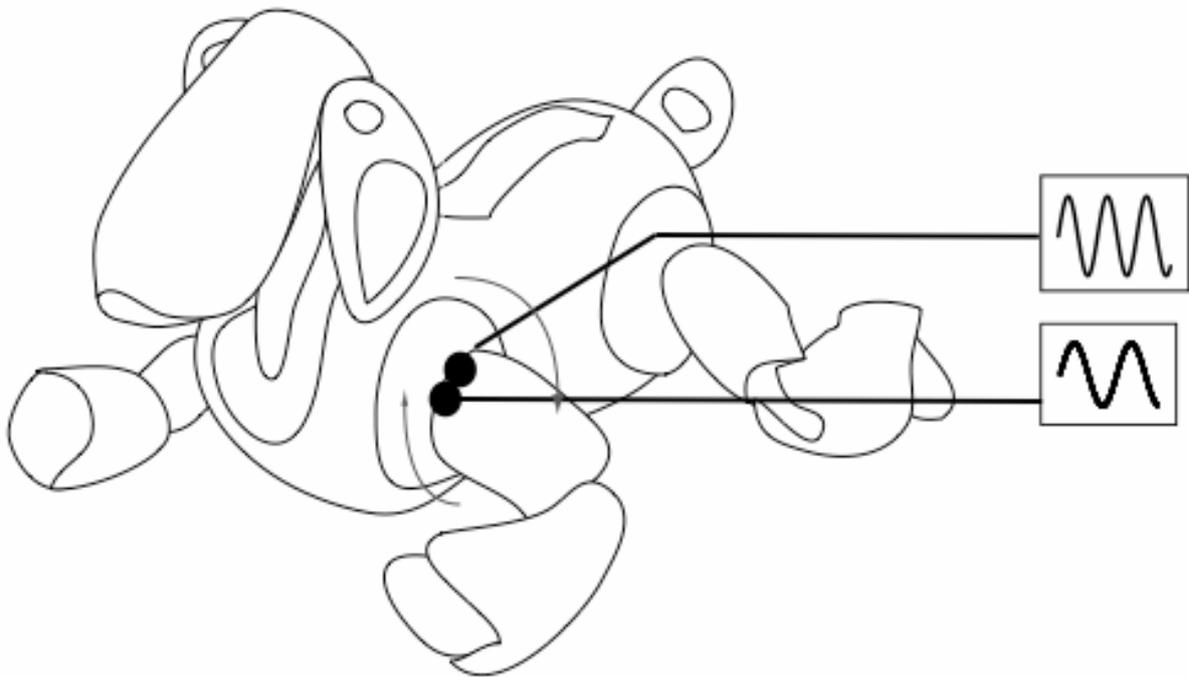


Figure 4 : Le robot quadrupède contrôle le mouvement des différents moteurs de ses jambes en appliquant des signaux sinusoïdaux. Deux signaux contrôlent le mouvement de chaque jambe et chaque signal est caractérisé par trois paramètres (amplitude, phase, fréquence). Son système de prédiction tente de prédire les effets de différents jeux de paramètres sur la manière dont l'image captée par sa caméra est modifiée, ce qui reflète indirectement le mouvement de son buste

Lorsque nous débutons une expérience de ce genre, le robot explore au hasard différents jeux de paramètres pendant les premières minutes. Il agite ses jambes de manière désordonnée. La majeure partie de ces mouvements a un effet très facilement prédictible : le robot ne bouge quasiment pas. Le robot, malgré son agitation, reste immobile. L'erreur en prédiction reste minimale : ces situations ne sont pas intéressantes pour le robot. Par

hasard, au bout d'une dizaine de minutes un mouvement amène en général le robot à effectuer un léger déplacement. Il découvre par exemple qu'une certaine combinaison de paramètres résulte en un léger recul. Cette situation nouvelle résulte d'abord en une augmentation de l'erreur en prédiction puis, au fur et à mesure que le robot a de nouvelles occasions d'effectuer des mouvements similaires, cette erreur commence à baisser : le robot a découvert une « niche de progrès ».

Dans l'heure qui suit, le robot va explorer les différentes manières de reculer. Au cours de cette exploration, il est probable qu'il découvre que certaines légères modifications de paramètres conduisent à effectuer ce qu'un observateur extérieur appellerait des mouvements de rotation : un nouvel ensemble de « niches de progrès » que le robot pourra exploiter quand les compétences liées à la marche en arrière auront été pour l'essentiel maîtrisées.

Il faut en général plus de trois heures pour que le robot découvre plusieurs ensembles de paramètres lui permettant de marcher en avant, en arrière, latéralement ou de tourner sur lui-même. A aucun moment, il n'a comme objectif d'apprendre à marcher. Guidé par la maximisation de la réduction de l'erreur en prédiction, il développe pourtant des compétences variées pour la locomotion. C'est d'ailleurs le caractère non spécifique de l'architecture qui permet cette variété. Un robot motivé pour se rapprocher d'un objet n'aurait par exemple sans doute pas appris à reculer ou à tourner sur lui-même.

Le fait que la marche en arrière se soit révélée dans cette trajectoire plus facile à découvrir que les autres n'était pas facile à prévoir. Etant donné la structure physique de ce robot et le type de sol sur lequel il était placé, les mouvements de recul ont été la première niche découverte. Pour savoir si cette niche constitue un « attracteur » récurrent pour ce type de trajectoire il faut mettre en place un programme d'expériences systématique semblable à celui que nous allons présenter dans le cadre de l'expérience qui suit.

4. Expérience 2: Découverte des objets et de la communication

Dans cette seconde série d'expériences, le robot est placé dans un environnement où plusieurs types d'objets sont présents. Un objet peut être mordu. Un objet peut-être poussé. Il y a également un autre robot qui imite les sons qu'émet le premier robot quand

celui-ci les produit dans sa direction. Le robot contrôle les paramètres d'un certain nombre de primitives motrices qui lui permettent un large éventail de mouvements. Il perçoit les conséquences de ces mouvements visuellement, de manière sonore et en utilisant ses capteurs de distance.



Figure 5 : Un robot explore son environnement sur un tapis d'éveil. Certains objets peuvent être frappés, d'autres peuvent être mordus. Un autre robot imite les sons qu'émet le premier robot quand celui-ci les produit dans sa direction.

Comme pour la découverte de la marche, le robot explore cet espace avec comme unique objectif de maximiser la réduction de ses erreurs en prédiction. Chaque expérience dure ici une dizaine d'heures et se caractérise par une structure remarquable (Figure 6). Après une première phase d'exploration aléatoire (Stade 0), le robot commence à effectuer de manière systématique des séries d'actions répétées (frapper, morde) (Stade 1), puis dans un second temps il découvre des zones de l'environnement (objets frappables, objets mordables) qui correspondent à des progrès en apprentissage (Stade 2). A ce stade

cependant, le robot ne fait pas d'association systématique entre une zone contenant un objet et l'action qui lui correspond. Il frappe l'objet à mordre, mord l'objet à frapper. Du point de vue de l'apprentissage il est cependant plus intéressant d'associer la bonne action au bon objet. La découverte de ces affordances (Gibson, 1977) a lieu quelques heures plus tard (Stade 3).

La production sonore, explorée dans un premier temps au même titre que les mouvements du corps, est rapidement abandonnée au profit de l'interaction avec les objets qui donne des résultats plus immédiats en terme de réduction d'erreur. En effet, lors des interactions sonores, le son émis par le robot est déformé quand l'autre robot l'imité. Cet effet est déterministe mais plus difficile à prédire, du moins dans un premier temps. Ce n'est qu'une fois l'interaction avec les objets maîtrisées que le robot commence à réémettre des sons (Stade 4). Il se consacre alors presque exclusivement à cette modalité.

Chaque expérience donne lieu à une trajectoire unique. Mais un même type de structuration se retrouve dans la majorité d'entre elles : le robot explore d'abord les tâches les plus simples pour se concentrer progressivement sur les situations les plus difficiles en termes de prédiction. Une trajectoire particulière n'est pas entièrement déterminée par les algorithmes qui contrôlent le robot. Elle n'est pas non plus la conséquence directe des opportunités présentes dans l'environnement. Elle résulte de l'interaction entre un système d'apprentissage générique, un corps robotique particulier et un environnement structuré. Grâce à cette approche, nous avons la possibilité d'étudier le rôle respectif de chacun de ces facteurs structurant dans l'organisation d'une séquence du développement.

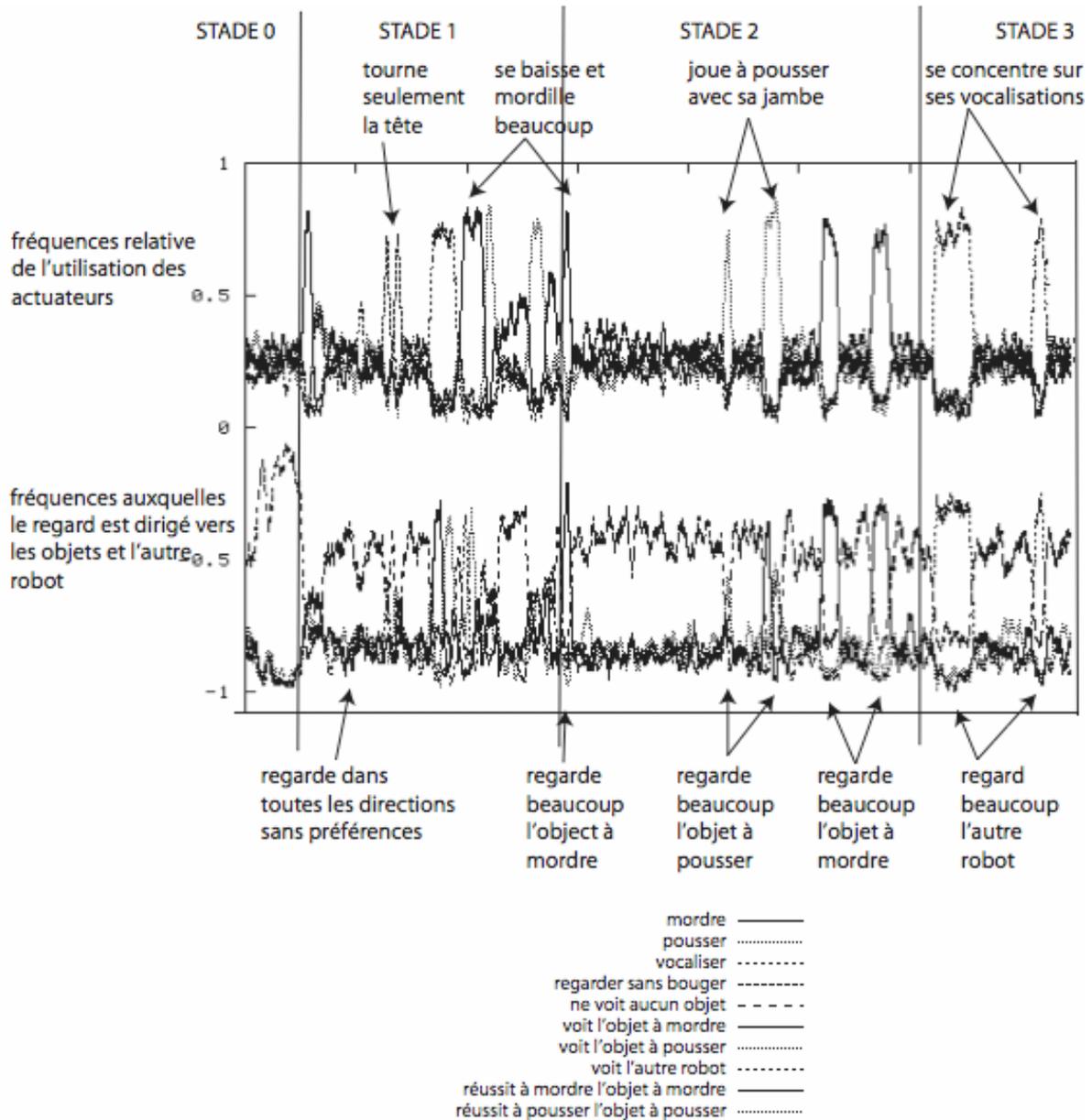


Figure 6 : Mesures des activités du robot durant les dix premières heures d'expériences. L'évolution de l'exploration sensorimotrice peut-être qualitative décomposée en différents stades (voir texte).

5. Discussion

Les recherches actuelles en neurosciences portent une attention particulière sur la manière dont le cerveau traite la nouveauté et les signaux d'erreur en prédiction (e.g. Schultz, et al 1997, Dayan et Belleine 2002, Kakade et Dayan 2001, O'Doherty, J. 2004).

Panksepp a regroupé un nombre important de données expérimentales et anatomiques suggérant l'existence d'un système responsable pour les comportements exploratoires (Panksepp 1998). Mais articuler de manière précise les liens entre ces caractéristiques neurophysiologiques et les processus de développement chez l'enfant reste à l'heure actuelle, essentiellement hors de portée.

Les systèmes d'apprentissage que nous utilisons ne sont pas directement inspirés par les données neurophysiologiques. Le corps de notre robot n'a pas grand chose à voir avec celui d'un enfant. Dans notre approche, il ne s'agit pas d'imiter l'homme en tout point. Au contraire, c'est en étudiant le rôle structurant des contraintes physiques et algorithmiques de la machine sur ses propres trajectoires de développement, que l'on peut en retour construire des explications pertinentes sur l'influence de facteurs comparables dans le développement du jeune enfant. En modifiant l'organisation de l'espace dans lequel le robot évolue ou le type robot utilisé, il devient possible d'étudier de manière systématique le rôle respectif de l'environnement, des contraintes anatomiques et morphologiques et des dynamiques motivationnelles dans le processus de développement artificiel. En répétant ces expériences de multiples fois, on peut rendre compte à la fois de la variabilité et de la robustesse de ce type de processus : chaque trajectoire est unique, fruit d'une histoire singulière avec un environnement spécifique mais toutes les trajectoires présentent des caractéristiques structurelles communes. Parce qu'elles permettent ainsi une démarche expérimentale d'un genre nouveau, ces expériences robotiques ouvrent la voie vers une meilleure compréhension de l'articulation entre contraintes spécifiques et mécanismes généraux d'apprentissage.

Références

- Barto, A., S. Singh, and N. Chentanez, (2004) Intrinsically motivated learning of hierarchical collections of skills, in *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)*, San Diego.
- Berlyne, D. (1960) *Conflict, Arousal and Curiosity*. McGraw-Hill.
- Csikszentmihalyi, M. (1991) *Flow-the psychology of optimal experience*. Harper Perennial.
- Dayan, P. and W. Belleine, (2002) Reward, motivation and reinforcement learning, *Neuron*, 36 :285–298,

- Deci, E. *Intrinsic Motivation*. New York: Plenum Press, 1976.
- Fedorov, V. *Theory of Optimal Experiment*. Academic Press, 1972.
- Gibson, E.J, (1988) Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge, *Ann. Rev. of Psychology*, 39 :247-263.
- Gibson, J. J. (1977) *The Theory of Affordances*. In R. E. Shaw & J. Bransford (eds.), *Perceiving, Acting, and Knowing*. Lawrence Erlbaum Associates, Hillsdale, NJ,
- X. Huang and J. Weng, (2002) “Novelty and reinforcement learning in the value system of developmental robots,” in *Proceedings of the 2nd international workshop on Epigenetic Robotics : Modeling cognitive development in robotic systems*, C. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius, Eds. LUCS 94, pp.47–55.
- Kakade, S. and P. Dayan (2002) Dopamine: Generalization and bonuses, *Neural Networks*, 15 :549–559.
- O’Doherty, J. (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging, *Current Opinion in Neurobiology* 14:769–776.
- Oudeyer P.-Y., Kaplan F. (2006) Discovering Communication, *Connection Science*, 18(2), pp. 189—206
- Oudeyer, P.-Y., Kaplan, F., Hafner, V.V. (in press), Intrinsic Motivation Systems for Autonomous Mental Development, *IEEE Transactions on Evolutionary Computation*.
- Panksepp, J. (1988) *Affective Neuroscience*, Oxford University Press.
- Schultz, W . and P. Dayan, and P. Montague, (1997) A neural substrate of prediction and reward, *Science*, 275 :1593–1599.
- Schmidhuber, J. (1991) Curious model-building control systems, in *Proceedings of International Joint Conference on Neural Networks*, pp. 1458–1463.
- Sutton, R. and A. Barto, (1998) Reinforcement learning: an introduction. Cambridge, MA.: MIT Press,
- Thrun, S. (1995) Exploration in active learning, in *Handbook of Brain Science and Neural Networks*, M. Arbib, Ed. Cambridge, MA: MIT Press.
- White, R. , (1959) Motivation reconsidered: The concept of competence, *Psychological review*, 66 :297–333.