L'apprentissage par renforcement comme résultat de la sélection

Samuel Delepoulle[*,**] — **Philippe Preux**[*] — **Jean-Claude Darcheville**[**]

[*] *Laboratoire d'Informatique du Littoral (LIL)*
UPRES-JE 2335
Université du Littoral Côte d'Opale
BP 719
F – 62228 Calais Cedex
{delepoulle,preux}@lil.univ-littoral.fr

[**] *Unité de Recherche sur l'évolution des Comportements et des Apprentissages (URECA)*
UPRES-EA 1059
Université de Lille 3
BP 149
F – 59653 Villeneuve d'Ascq Cedex
darcheville@univ-lille3.fr

RÉSUMÉ. *Dans cet article, en utilisant des simulations, nous montrons que l'apprentissage par renforcement peut résulter de la sélection naturelle. L'apprentissage par renforcement est un élément essentiel de la dynamique du comportement animal. Pour cela, nous nous appuyons sur des hypothèses issues de l'étude du comportement animal. Toujours en utilisant des simulations informatiques, nous montrons ensuite que la capacité d'apprendre par renforcement confère la possibilité de mettre en place des interactions riches entre plusieurs agents en les mettant dans des situations bien étudiées en psychologie sociale.*

ABSTRACT. *In this paper, using computer simulations, we show that the ability to perform reinforcement learning may result from natural selection. Reinforcement learning is an essential element of the dynamics of animal behavior. To this end, we ground our work on hypothesis originating from the study of animal behavior. Then, again using computer simulations, we show that the ability to learn by reinforcement may yield to rich interactions in a group of agents that we simulate in situations studieddrawn from social psychology.*

MOTS-CLÉS : *apprentissage, sélection naturelle, effet Baldwin, simulation*

KEYWORDS: *learning, natural selection, Baldwin effect, simulation*

ECA – 1/2001. Apprentissage et évolution. Volume - n° /, pages 9 à 30

1. Introduction

Depuis ses origines, l'informatique tente de s'inspirer des systèmes naturels et de certains éléments des systèmes vivants en particulier. Ainsi, très rapidement se sont développés des travaux sur les neurones et les réseaux de neurones formels bâtis sur le modèle de neurones proposé par Pitts et McCulloch [PIT 47]. De même, des chercheurs ont rapidement tenté de comprendre l'évolution des espèces [HOL 61] pour en retirer des algorithmes [FOG 66, REC 73, HOL 75] largement popularisés dans les années 1980/90 sous le terme générique d'algorithmes évolutionnaires. On peut aussi citer ici les travaux sur l'algorithme de Métropolis [MET 53] et plus récemment, les algorithmes basés sur les simulations de colonies d'insectes sociaux [BON 99], mais aussi les algorithmes de renforcement [SUT 98]. Source d'inspirations, la biologie (comme « étude du vivant » au sens large) peut bénéficier en retour de la simulation informatique de ces systèmes : lors des simulations, certaines dynamiques sont observées qui rappellent celles observées dans le monde du vivant ; ces simulations peuvent également être utilisées comme modèles de processus biologiques et aider à leur validation, à leur réfutation, ou, au moins, apporter des arguments aux uns et aux autres pour débattre. Les systèmes ainsi étudiés sont complexes, c'est-à-dire qu'ils comprennent plusieurs agents en interaction, ces interactions donnent lieu à des dynamiques non linéaires difficilement appréhendables intuitivement *a priori* (ce qui peut faire croire hâtivement à l'émergence de nouvelles propriétés [FOR 91]).

La mise à jour de l'ensemble des processus en interaction et leur compréhension demeure un enjeu-clé pour la compréhension de notre monde que seule une approche pluridisciplinaire et transdisciplinaire pourra accomplir. Nous pouvons tenter de préciser différents niveaux de processus (voir la figure 1). Au niveau le plus fondamental, on trouve le processus d'évolution génétique, moteur de l'évolution des espèces vivantes. Au cours de la phase de développement, le génome, acompagné de son armada de molécules, engendre différents types de cellules qui donneront naissance aux organes ou à différents composants de l'organisme : cellule nerveuse, cellule musculaire, cellule du système immunitaire...Cerveau, muscles, os, tendons, peau...sont ainsi formés, capables d'activités innées, ou « réflexes ». Les organismes formés de cette manière sont capables d'apprendre au cours de leur vie de nouveaux comportements : voir, regarder, saisir un objet, marcher, faire du vélo (pour certaines espèces), utiliser un crayon (même remarque)...*via* un apprentissage dit « opérant » qui consiste à sélectionner le bon (du moins, un pas trop mauvais) comportement à émettre dans un contexte perceptif donné (cf. *infra*). Cet apprentissage s'appuie sur la plasticité de notre cerveau et de nos organes. Acquis au cours de la vie, ces comportements peuvent devenir des réflexes lorsqu'ils sont maîtrisés. En interaction avec d'autres organismes de son espèce et d'autres espèces, capables par apprentissage opérant d'apprendre à émettre de nouveaux comportements, un organisme peut alors apprendre des comportements *via* d'autres organismes en les mimant, en essayant de suivre des conseils, ou en respectant des « lois ». S'il peut sembler anthropomorphique, ce discours s'applique très bien aux animaux vivant en société ou, au moins, qui élèvent leurs petits [WIL 75]. A cet enchaînement de processus de bas en haut (du chimique au social), s'ajoutent

des processus allant du haut vers le bas. On pense aujourd'hui que la spécialisation morphologique et comportementale observée dans de nombreuses espèces d'insectes sociaux (perte au cours de l'évolution de l'espèce de certains organes, hypertrophie d'autres, stérilité pour certains de ses membres...) résulte de la rétroaction de l'organisation de la société sur les génomes de l'espèce. Expérimentalement, Waddington a été le premier à montrer sur des drosophiles que la plasticité du développement morphologique peut influencer la sélection génétique pour « canaliser[1] » une certaine capacité d'adaptation et amener une « assimilation génétique[2] » [WAD 53, WAD 56]. Cette rétro-action est un exemple de l'effet Baldwin du nom de l'un de ceux qui l'ont proposé, indépendamment en 1896, par Lloyd Morgan [MOR 96], Osborn [OSB 96] et Baldwin [BAL 96]. Ces auteurs ont proposé que la capacité d'adaptation au cours de la vie[3] d'un nouveau comportement ou d'un trait morphologique peut influencer l'évolution génétique jusqu'à ce que cette nouvelle aptitude devienne innée, codée dans le génome.

L'effet Baldwin a été étudié par simulation informatique, initialement par Hinton et Nowlan en 1987 [HIN 87]. Ceux-ci ont montré qu'une population d'organismes capable de réaliser un apprentissage assez rudimentaire placée dans une situation de problème de type « aiguille dans une meule de foin » évolue rapidement vers une population dont tous les individus résolvent de manière innée le problème. Plus précisément, la capacité à apprendre d'un individu est codée dans son génome d'une manière non déterministe : le génome d'un individu lui confère une certaine probabilité de pouvoir résoudre le problème ; la fitness des individus de la population est déterminée par leur aptitude à trouver plus ou moins rapidement l'aiguille dans la botte de foin ; alors qu'une population d'individus incapables d'apprendre stagne, une population d'individus pouvant apprendre évolue donc vers des individus résolvant le problème de manière innée. La raison en est que l'évolution génétique, lente puisqu'elle agit au rythme des générations, est aidée par l'exploration réalisée au cours de la vie dans le cas des individus ayant la capacité d'apprendre ; formellement, cela s'explique par un problème de combinatoire. En quelque sorte, l'apprentissage catalyse l'évolution génétique et l'accélère. Le point important ici est que la capacité d'adaptation au cours de la vie rétroagit positivement sur l'évolution génétique de l'espèce.

Depuis Hinton et Nowlan, l'interaction entre les processus d'apprentissage et d'évolution est beaucoup étudiée. Elle a montré son intérêt dans le domaine de l'optimisation sous la forme d'algorithmes dits « hybrides » combinant un algorithme évolutionnaire et un algorithme de recherche. En parallèle, plusieurs recherches consistent à étudier la dynamique de cette interaction (voir le « connexionisme génétique » de [CHA 90], [ACK 92, FLO 93, MIT 96a, LIT 96, PAR 96], et [FLO 99, URZ 00] pour des états de l'art récents sur le sujet). Divers auteurs ont également souhaité enrichir cette interaction avec des aspects culturels (voir par exemple [BEL 90]).

1. Mot utilisé par Waddington.
2. *Idem.*
3. Adaptation (au cours de la vie) et apprentissage sont ici des synonymes.

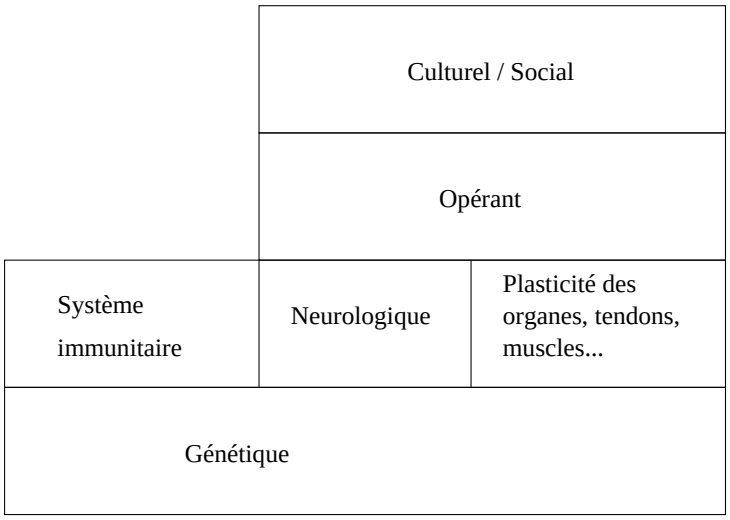

Figure 1. *Ce schéma représente quelques processus importants d'apprentissage. Au niveau le plus fondamental est situé l'apprentissage génétique qui se déroule au long des générations, par sélection naturelle ; sur cette couche s'appuient plusieurs processus d'apprentissage, ayant des fondements innés mais se déroulant tout au long de la vie dont nous citons trois exemples : le système immunitaire qui effectue un apprentissage des agents infectieux, l'apprentissage neurologique pour le développement des structures corticales, et un ensemble de composants organiques des organismes vivants ; sur les deux derniers élements de ce niveau s'appuie la brique de base de l'apprentissage (au sens classique du terme) qualifiée techniquement d'apprentissage opérant ; enfin, au niveau supérieur, on trouve les apprentissages liés aux interactions entre organismes dans leur société, entre parents et petits…Ce schéma mélange des processus ayant lieu à des échelles temporelles différentes, agissant sur des entités différentes (population, individus, organes d'un individu…). Les interactions entre niveaux ont lieu de bas en haut, mais aussi du haut vers le bas (effet Baldwin entres-autres)*

Parmi les multiples questions à étudier, la capacité à apprendre elle-même devrait être expliquée par la sélection naturelle. Cela signifie que si la sélection naturelle est invoquée comme processus de base de l'évolution, l'apprentissage opérant doit avoir été sélectionné au cours de l'évolution.

Au niveau le plus simple, l'apprentissage par renforcement peut être défini comme la capacité d'un animal à modifier son comportement en fonction des stimuli qu'il reçoit de son environnement. Ceci a été modélisé par Thorndike en tant que « loi de l'effet ». La loi de l'effet stipule que la fréquence d'émission de certains comportements augmente quand leur émission a été suivie de conséquences favorables dans le passé [THO 98, THO 11]. La loi de l'effet a été étudiée expérimentalement dans de

très nombreux travaux et par une large communauté de recherche. Skinner a proposé le principe de la sélection du comportement par ses conséquences [SKI 38, SKI 81] qui repose sur les mêmes idées, bien que le cadre conceptuel ait évolué depuis Thorndike [CHA 99]. Ce principe demeure d'actualité pour comprendre l'évolution des comportements complexes [STA 00]. La loi de l'effet est un bon exemple de la difficulté de formalisation d'un modèle évoquée plus haut. En effet, de multiples modèles ont été proposés [SUT 98] mais sont encore loin de modéliser parfaitement, et de manière non *ad hoc*, certains processus élémentaires observés dans le vivant. Ce point a été étudié dans [DEL 00c].

Cette capacité à adapter le comportement en fonction de ses conséquences a été mis en évidence par les procédures d'apprentissage opérant. L'apprentissage opérant, qui est une forme d'apprentissage par renforcement, peut-être vu comme un niveau élémentaire d'adaptation du comportement. Dans le cas de l'apprentissage par renforcement, une stimulation particulière de l'environnement (le renforçateur) suit l'emmission d'un comportement et en augmente la probabilité d'apparition.

Le caractère adaptatif du comportement a été étudié par le biais de nombreuses procédures qualitatives et quantitatives (conditionnement classique, conditionnement opérant, apprentissage discriminatif, modelage du comportement...). Signalons également que l'apprentissage par renforcement est observée chez la quasi-totalité des espèces du règne animal. On peut vraisemblablement expliquer cela par le fait que l'apprentissage par renforcement a été sélectionné par l'environnement. A l'heure actuelle, aucune étude empirique n'a pu établir ce fait pour les organismes naturels. Nous proposons d'examiner si cette propriété peut apparaître chez des agents artificiels soumis à une sélection génétique. Les principales propriétés de l'apprentissage opérant étant établies, il est donc envisageable d'en réaliser des modélisations et des simulations informatiques.

Dans la suite, nous commençons par préciser le modèle utilisé pour les agents et les processus qui simulent l'évolution naturelle et l'apprentissage. Ensuite, nous présentons les tâches auxquelles ces agents sont confrontés et qui constituent l'environnement dans lequel ils « vivent ». Après quoi, nous présentons les résultats de simulations. Nous terminons par une discussion sur ce que ce travail apporte et ses perspectives.

2. Le modèle

Dans cette section, nous décrivons le modèle, c'est-à-dire, les agents adaptatifs ainsi que les processus de l'évolution génétique et des comportements. L'évolution génétique modélise la sélection naturelle agissant à l'échelle des générations sur une population d'individus ; l'évolution des comportements modélise quant à elle l'adaptation des comportements (ou « apprentissage ») agissant à l'échelle de sa vie sur un individu. Le phénotype d'un agent est obtenu par expression de son code génétique. Celui-ci code essentiellement un réseau de neurones (le phénotype de l'agent) lequel

contrôle le comportement de l'agent durant sa « vie ». Le génome code des grandeurs liées au comportement dynamique du réseau plutôt que des caractéristiques statiques comme sa topologie. Dans un but de plausibilité biologique, ce réseau implante les propriétés suivantes :

– le réseau est non-supervisé : il n'y a jamais de présentation du « bon » comportement qui aurait du être émis, ni même d'indication concernant le fait qu'un meilleur comportement aurait pu être émis ou non. Il n'y a pas non plus d'indication dans l'environnement sur le comportement attendu, sur les associations stimulus/réponse qui doivent être apprises ;

– le réseau simule un fonctionnement parallèle. Chaque neurone est activé aléatoirement dans le temps ;

– les connexions du réseau se font dans deux sens. Les connexions dans le sens entrée-sortie sont appelées « entrantes » alors que celles qui sont dans le sens sortie-entrée sont dénommées « ré-entrantes ». De cette façon, le système « perçoit » les comportements qu'il émet (proprioception) ;

– l'activité des neurones à un moment donné est déterminée par leur propre activité dans le passé. Chaque neurone possède donc une boucle de rétro-action qui permet le maintien dans le temps de son activité, ce qui réalise une sorte de mémoire à long terme.

La sélection naturelle est simulée par l'utilisation d'un algorithme génétique.

2.1. *Les agents*

Chaque agent est constitué de trois élements (voir figure 2) :

1) un ensemble de N entrées sensorielles (ES) qui lui permettent de percevoir son environnement,

2) un ensemble de N unités comportementales (UC) qui lui permettent d'effectuer une action sur son environnement,

3) un réseau de neurones qui contrôle ses activités de façon adaptative pendant le cours de sa vie.

Le réseau de neurones d'un agent est constitué de C couches, comprenant chacune N neurones. Les entrées ES reçoivent un stimulus binaire venant de l'environnement. D'une manière générale, nous dénommons « unité » à la fois les entrées sensorielles, les unités comportementales et les neurones. Chaque neurone reçoit la sortie des N unités de la couche précédente (connexions entrantes) et la sortie des N unités de la couche suivante (connexions de ré-entrance) ; donc, chaque neurone reçoit $2N$ entrées. Grâce à ces connexions, le réseau de neurones d'un agent perçoit ses propres comportements puisque les unités comportementales (UC) ont une influence sur la couche de sortie du réseau.

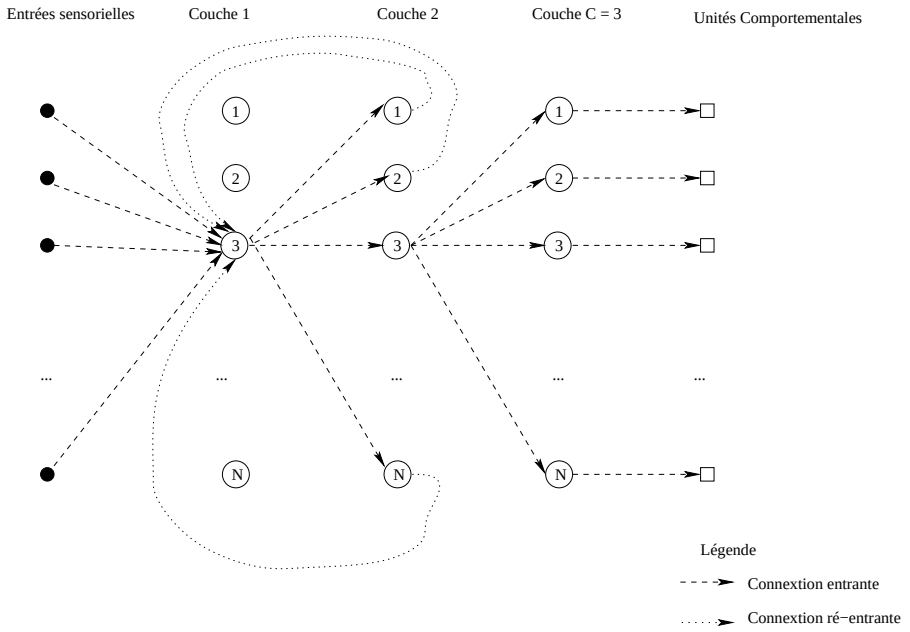

Figure 2. *Structure des agents. Pour simplifier le schéma, toutes les connexions ne sont pas représentées. Sur le neurone 3 de la couche 1, on peut voir 4 des N entrées provenant des entrées sensorielles et 3 des N connexions de réentrance qui viennent des neurones de la couche 2. Chaque neurone reçoit des informations des neurones de la couche précédente (des entrées sensorielles pour la couche 1) et des informations des neurones de la couche suivante (des unités comportementales pour la couche 3)*

Chaque unité comportementale est connectée à un neurone de la dernière couche du réseau par une relation un à un. A chaque pas de temps, une et une seule unité comportementale est active, celle associée au neurone qui a le potentiel le plus élevé, selon une règle *winner takes all* ; si le potentiel le plus élevé est le même pour plusieurs neurones, un tirage aléatoire est effectué parmi eux pour déterminer l'UC qui est active.

Dans cet article, C et N sont toujours fixés respectivement à 3 et 10. Donc il y a une couche de neurones d'entrée, une couche de sortie et une couche cachée. Les caractéristiques du réseau de neurones sont codées dans un génome comme le montre la figure 3. Ce génome détermine les caractéristiques des neurones et des connexions. La réponse de chaque neurone est caractérisée par une valeur booléenne qui indique si le neurone est actif ou non et par 6 nombres réels : $\alpha, \beta, \gamma, a, b \in [-100, 100], \epsilon \in [-1, 1]$. Ces 6 paramètres déterminent la réponse du neurone en fonction de l'activité des neurones environnants et de sa propre activité au pas de temps précédent (comme cela sera précisé par la suite).

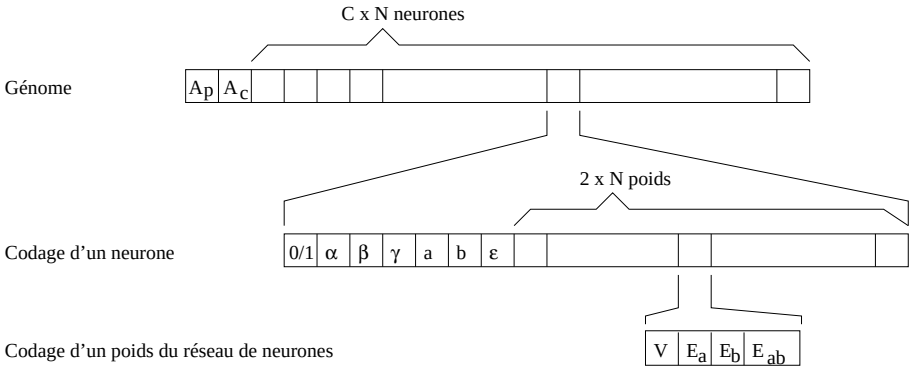

Figure 3. *Codage génétique du phénotype d'un agent : un réseau de neurones représenté par un ensemble de $C \times N$ neurones. En plus d'un bit d'activation et de 6 valeurs réelles, chaque neurone est caractérisé par $2 \times N$ poids, chacun l'étant par 4 valeurs élémentaires. L'ensemble est précédé par les valeurs A_p et A_c. L'influence de chacun des paramètres est exposée dans le texte*

Dans la mesure où chaque neurone est connecté en sortie à chacun des neurones des deux couches voisines, chaque neurone est également caractérisé par $2 \times N$ poids à valeur réelle. Chaque poids est déterminé par un quadruplet (V, E_a, E_b, E_{ab}), où V est la valeur initiale, E_a, E_b et E_{ab} contrôlent l'évolution de la valeur au cours de l'apprentissage. La valeur de ces 4 paramètres est comprise dans l'intervalle $[-1, 1]$. Enfin, l'ensemble du réseau est caractérisé par deux nombres entiers A_c and A_p. Ces valeurs déterminent le nombre de fois où l'algorithme d'activation et l'algorithme d'apprentissage sont exécutés.

On voit donc que le génome ne code pas les poids synaptiques d'un réseau mais des informations permettant de calculer ce poids et décrivant l'évolution de sa valeur au cours de l'apprentissage. Nous nous situons donc entre l'utilisation d'un codage où le génome code directement le poids des connexions synaptiques et un codage décrivant des règles d'évolution de ces poids comme dans [FLO 96] ; dans notre cas, les règles d'évolution sont fixées, mais leurs paramètres sont codés dans le génome. Nous nous rapprochons ainsi du connexionisme génétique de Chalmers [CHA 90] ayant pour objectif de coder des propriétés dynamiques plutôt que statiques dans le génome décrivant un réseau de neurones. Par rapport à ce travail, notre approche s'en distingue sur différents points : nos réseaux sont plus complexes (ils sont multi-couches et réentrants) ; le génome que nous utilisons est beaucoup plus long et plus riche (les valeurs sont codées par un nombre réel alors que Chalmers les code sur 3 à 5 bits) ; nous n'utilisons pas d'apprentissage supervisé dans nos réseaux ; nous nous appuyons sur des hypothèses liées à l'étude du comportement animal ; l'environnement dans lequel évolue nos agents est dynamique contrairement à celui utilisé par Chalmers. Notons bien que dans notre travail, l'architecture du réseau est fixe et qu'elle n'est pas codée

dans le génome, contrairement à certaines approches où une phase de développement transforme le génome en un réseau de neurones (voir [GRU 92, MIG 96, KOD 98]).

2.2. *Évolution génétique*

Le processus d'évolution est simulé en utilisant un algorithme génétique qui opère sur le génome décrit précédemment. Rappelons très brièvement qu'un algorithme génétique agit itérativement sur une population de génomes ; à chaque itération, la fitness de chacun des génomes est évaluée en fonction de la performance du phénotype qu'il exprime (de manière plus ou moins déterministe) ; en fonction de cette fitness, certains individus produisent de nouveaux individus en combinant leur génome avec celui d'autres individus de la population (opération de recombinaison ou cross-over) et en modifiant aléatoirement le génome résultant (mutation) ; recombinaison et mutation constituent les opérateurs génétiques de l'algorithme génétique ; pour plus de détails généraux sur ces algorithmes, on consultera par exemple [MIT 96b].

Plus précisément, à chaque génération de l'algorithme génétique, la fitness de chacun des individus de la population est évaluée en le faisant résoudre une tâche : cette partie simule l'apprentissage au cours de la vie de l'agent et a pour objectif d'évaluer sa capacité d'apprentissage. La fitness d'un agent est d'autant plus grande qu'il obtient de bonnes performances sur la tâche. Nous décrivons maintenant la phase de reproduction et les opérateurs génétiques.

2.2.1. *La reproduction*

Pour constituer la population des descendants, les deux individus ayant la fitness la plus basse sont éliminés de la population. Ils sont remplacés par deux descendants des deux agents ayant la fitness la plus importante dans la population. On utilise donc un schéma « steady-state » où le meilleur individu de la population est recombiné avec un autre individu pour produire deux nouveaux individus qui remplacent les deux moins bons individus de la population. Pour former les deux enfants, nous utilisons 6 opérateurs de variation : une recombinaison et 5 types de mutation.

2.2.2. *La recombinaison*

Un crossover un point est utilisé sur deux génomes pour produire un descendant. Le crossover un point ne peut couper qu'entre deux neurones. L'individu reçoit les deux paramètres A_c et A_p de l'un des parents aléatoirement.

2.2.3. *La mutation*

Cinq opérateurs de mutation sont utilisés. Ce nombre important d'opérateurs de mutation est dû au fait que chacun agit à un certain niveau du génome : poids, neurone, expression, réponses et apprentissage. D'un point de vue informatique, cela correspond à des types de données différents, donc à des actions différentes à réaliser

pour les modifier ; dans un génome réel, ces mutations correspondent à des erreurs de recopie frappant différentes zones du génome.

La première mutation agit sur les poids. Elle consiste à choisir aléatoirement un poids dans l'ensemble du génome et à modifier sa valeur dans une fourchette de 10 % de la valeur initiale. Ceci provoque donc une mutation qui a un faible effet. La probabilité que cette mutation apparaisse est notée μ_w. Elle peut être assez élevée dans la mesure où ses effets ne sont pas très destructeurs.

La seconde mutation modifie le neurone et consiste à réinitialiser aléatoirement toutes les caractéristiques du neurone. Dans ce cas, toutes les valeurs $\alpha, \beta, \gamma, a, b$ et ϵ sont réinitialisées. Il en va de même des valeurs qui correspondent au $2 \times N$ poids de ce neurone. La probabilité de cette mutation est notée μ_n ; ses effets sur l'activité du réseau sont nettement plus importants. Donc nous utilisons un taux plus faible pour μ_n.

La troisième mutation concerne l'expression d'un neurone et modifie simplement un bit d'activité d'un neurone dans le génome. Le changement de ce bit peut avoir des conséquences importantes sur l'activité du réseau. Quand il est inactif, un neurone peut « voyager » de génération en génération sans s'exprimer. Des mutations peuvent alors intervenir sans affecter le comportement de l'agent (mutation neutre). Quand il est réactivé, l'accumulation de ces mutations neutres peut modifier fortement l'activité du réseau et donc sa capacité d'adaptation. Elle est appliquée avec la probabilité μ_e.

Les deux dernières mutations concernent les paramètres A_c et A_p. Avec une probabilité μ_p, chacune de ces deux variables peut être modifiée indépendamment. Leur mutation change leur valeur de ± 10 unités au maximum.

2.3. *Evolution des comportements*

Dans cette section, l'apprentissage d'un agent au cours de sa « vie » est décrit. Avant cela, nous décrivons la réaction d'un agent aux stimuli pour produire un comportement. L'algorithme principal consiste à faire répondre puis apprendre chacun des agents pendant toute la durée de leur vie. Répondre consiste à choisir aléatoirement successivement A_r neurones et à mettre à jour leur activité ; apprendre consiste à choisir aléatoirement A_p connexions et à mettre à jour leur poids.

2.3.1. *Activation*

Pour s'approcher d'une activité parallèle des neurones, les neurones d'un agent sont activés comme suit. Itérativement, A_c neurones sont sélectionnés au hasard, en laissant la possibilité qu'un neurone soit activé plusieurs fois pendant une même exécution de « Répondre ». Soient $l \in U(1, C)$ [4] et $n \in (1, N)$ représentant respectivement la couche et le numéro du neurone à activer, ce neurone doit être fonction-

4. $U(a, b)$ désigne le tirage d'un entier pseudo-aléatoire dans $[a, b]$ selon une loi de probabilité uniforme.

nel (bit d'activité sur « on »). S'il n'est pas fonctionnel, son activité est considérée comme nulle. Notons $A_t(l,n)$ l'activation de ce neurone au temps t. $A_{t+1}(l,n)$ est écrit comme une fonction de l'activation courante $A_t(l,n)$, de la somme pondérée des entrées $Se_t(l,n)$, de la somme pondérée des réentrances $Sr_t(l,n)$ et d'un facteur aléatoire réel $h_t \in U(-1,1)$ qui joue le rôle de bruit blanc. Alors, l'étape « Mettre à jour son activité » peut être écrite comme suit :

$$A_{t+1}(l,n) = f(\alpha_{ln}.Se_t(l,n) + \beta_{ln}.Sr_t(l,n) + \gamma.A_t(l,n) + \epsilon_{ln}.h_t)$$

avec

$$\begin{cases} Se_t(l,n) = \sum_{k=1}^{N} V_t^e(k.ln) \times A_t(l-1,k) \\ Sr_t(l,n) = \sum_{k=1}^{N} V_t^r(k.ln) \times A_t(l+1,k) \end{cases}$$

où $V_t^e(k.ln)$ est le poids au temps t de la k^{e} connexion entrante du neurone l,n et $V_t^r(k.ln)$ est le poids au temps t de la k^{e} connexion de réentrance du même neurone. La fonction $f(x)$ est linéaire par morceaux. Elle est déterminée par les constantes a_{ln} et b_{ln} :

– si $a_{ln} \neq b_{ln}$, alors $g(x) = 2(x - a_{ln})(a_{ln} - b_{ln}) - 1$

et

$$\begin{cases} f(x) = -1 & \text{si } g(x) \leq -1 \\ f(x) = g(x) & \text{si } -1 < g(x) < +1 \\ f(x) = +1 & \text{si } g(x) \geq +1 \end{cases}$$

– si $a_{ln} = b_{ln}$, alors

$$\begin{cases} f(x) = -1 & \text{si } x < a_{ln} \\ f(x) = +1 & \text{si } x \geq a_{ln} \end{cases}$$

L'activation de tous les neurones qui ne sont pas mis à jour reste inchangée. Finalement, si un neurone est « inactivé », son potentiel reste toujours nul.

2.3.2. *Apprentissage*

Comme on l'a dit plus haut, l'apprentissage n'est pas déterminé par les gènes mais les variables qui sont codées génétiquement interviennent dans la modification des réponses du réseau au cours du temps. En fait, l'apprentissage n'est pas entièrement prédéterminé génétiquement mais reste au contraire sous l'influence de l'environnement dans une large part. Nous appelons « Apprentissage » les modifications de l'activation du réseau en fonction des stimulations qu'il a reçues de l'environnement. Ceci consiste en une modification des poids du réseau. Ces modifications ne sont pas purement déterministes : une connexion du réseau est choisie aléatoirement et son poids est modifié en fonction des neurones auxquels elle est connectée. Là encore, une connexion peut être sélectionnée plusieurs fois au cours d'une seule exécution de « apprendre ».

Plus précisément, soit l,n et c, tous trois tirés dans $U(1,N)$, et $r \in U(1,2)$.

Si $r = 1$, le poids d'une connexion d'entrée est mise à jour comme suit :

$$V_{t+1}^e(c.ln) = V_t^e(c.ln) + E_a \times A_t(l,n) + E_b \times A_t(l-1,c) + E_{ab} \times A_t(l,n) \times A_t(l-1,c)$$

Si $r = 2$, le poids d'une connexion de ré-entrance est mise à jour comme suit :

$$V^r_{t+1}(c.ln) = V^r_t(c.ln) + E_a \times A_t(l,n) + E_b \times A_t(l+1,c) + E_{ab} \times A_t(l,n) \times A_t(l+1,c)$$

On peut préciser que l'apprentissage réalisé ici est un apprentissage non supervisé. D'une part, le comportement à émettre pour une entrée donnée n'est pas présenté au réseau pour qu'il corrige ses erreurs comme dans un apprentissage supervisé. D'autre part, la valeur récompense n'est pas une récompense pour le réseau lui-même ; cette valeur est utilisée pour déterminer la fitness de l'agent et donc sa probabilité de survie à la génération suivante, mais pas pour que le réseau tente de corriger ses comportements inadéquats au cours de sa vie, comme dans un apprentissage par renforcement.

Ceci termine la présentation du modèle.

Un lecteur non habitué pourrait être surpris par la simplicité du modèle par rapport au système réel supposé être modélisé. Pour le rassurer, nous indiquerons que les réseaux de neurones et les algorithmes génétiques sont classiquement utilisés dans ce contexte ; de plus, certaines caractéristiques de notre modèle ont été mises en place pour ajouter à la vraisemblance. Il demeure clair que notre modèle reste très simple, même s'il est déjà relativement complexe. Néanmoins, modéliser, c'est construire une métaphore de l'objet étudié qui en est une simplification, tout en gardant son essence.

3. Simulation

Les agents ont été soumis à trois tâches. Dans chaque cas, la fonction fitness est directement reliée au comportement des agents. Dans un premier temps, nous décrivons les tâches auxquelles les agents sont soumis. Ensuite, nous décrivons les résultats des simulations.

3.1. *Les tâches*

Nous décrivons trois conditions dans lesquelles les agents ont évolué. Ces conditions qui reprennent des situations de psychologie expérimentale sont décrites sous le nom de « tâche de discrimination », de « contrôle mutuel du destin » et enfin, nous introduisons une tâche dérivée le « contrôle mutuel du destin avec sélection du comportement ».

3.1.1. *Tâche de discrimination*

L'objectif de cette procédure est de sélectionner des agents qui sont capables de réaliser un apprentissage opérant, c'est-à-dire des agents capables d'apprendre à émettre des comportements qui ont été suivis par des conséquences favorables dans le passé. Rappelons que l'apprentissage opérant est un fondemenet de l'analyse expérimentale du comportement et qu'il modélise la capacité à apprendre et à adaptater son comportement au cours de l'existence. Plus précisément, cette tâche consiste à discriminer deux stimuli S1 et S2. En présence de S1, le comportement de l'agent doit être

B1 alors qu'en présence de S2, le comportement attendu est B2. S1 et S2 sont envoyés sur deux entrées sensorielles différentes (ES). B1 et B2 correspondent au comportement observé, c'est-à-dire à l'activation de deux entités différentes de la couche UC. Quand le comportement attendu est émis, un stimulus est envoyé sur l'une des ES, différente des ES activées par S1 et S2. Ce stimulus joue le rôle d'une conséquence positive (renforcement). Notons que cette conséquence positive n'est pas considérée en tant que telle par l'agent ; celui-ci reçoit simplement un stimulus sur l'une de ses ES, rien de plus.

L'objectif est d'obtenir un réseau capable d'apprendre une relation ; pour cela, il faut éviter que les relations S1-B1 et S2-B2 ne deviennent « câblées » dans l'agent. La tâche est composée de sessions. Au cours de chaque session, une association parmi S1-B1/S2-B2 et S1-B2/S2-B1 est choisie arbitrairement comme étant l'association qui procure le renforcement. Aucun signal n'est donné à l'agent concernant l'association active pendant la session : il doit l'apprendre par essai-erreur. Au début de chaque session, une association est déterminée aléatoirement avec une probabilité de 0.5, seule cette association étant renforcée. Grâce à cela, les agents doivent démontrer une capacité d'adaptation, c'est-à-dire qu'ils doivent être capables d'apprendre la bonne association stimulus-comportement ; plus généralement, ils doivent apprendre à modifier leur comportement au cours de la durée de leur « vie » en fonction des stimuli qu'ils reçoivent de l'environnement. Chaque session est composée de 1 000 cycles de présentation des stimuli suivis d'une réponse de l'agent. Chaque stimulus S1 ou S2 est émis avec une probabilité de 0.5 pendant la session.

La population est constituée de 10 agents. La fonction fitness est définie par le nombre cumulé de renforcements reçus pendant 10 sessions. La valeur maximale de la fonction fitness est donc 10 000. Initialement, la population est constituée d'agents dont les caractéristiques sont aléatoires. Les valeurs μ_w, μ_n, μ_e et μ_p sont respectivement fixés à 0.05, 0.01, 0.05 et 1.0. Initialement, les valeurs A_p et A_c sont tirées aléatoirement dans l'intervalle $[1, 100]$.

3.1.2. *Contrôle mutuel du destin*

La procédure de contrôle mutuel du destin (CMD) provient d'études de psychologie sociale et modélise une situation de coopération. Elle a été introduite dans les années 50 [SID 56]. L'idée est de confronter deux agents A et B qui peuvent choisir entre deux comportements B1 et B2. Les comportements d'un agent n'ont de conséquence que pour l'autre agent :

– si le comportement de A est B1, alors B gagne un point ;

– si le comportement de A est B2, alors B perd un point ;

– si le comportement de B est B1, alors A gagne un point ;

– si le comportement de B est B2, alors A perd un point.

Donc, pour chaque agent, son comportement n'a aucune conséquence pour lui-même. Il peut uniquement jouer sur le gain de l'autre (c'est dans ce sens que chacun contrôle le « destin » de l'autre). Cette situation conduit à une dynamique complexe que nous

avons discutée dans [DEL 00a, DEL 00b, DEL 01a]. Pour la suite, nous plaçons les agents dans cette situation. Comme pour la tâche précédente, le but est de sélectionner des agents adaptatifs (et non des agents qui répondraient systématiquement B1). Pour cela chaque agent effectue à nouveau 10 sessions. Là encore, dans la moitié des cas, les gains sont inversés : si A (resp. B) choisit B2, alors B (resp. A) reçoit un point et si A (resp. B) choisit B1, alors B (resp. A) perd un point.

Dans la mesure où les gains sont contrôlés exclusivement par le comportement de l'autre agent, nous confrontons chaque agent à un « clone » de lui-même. De cette manière, au cours de l'évolution génétique il n'y a pas de risque que des agents incapables de réaliser la tâche éliminent ceux qui en sont capables. Cette procédure permet d'évaluer la capacité de l'algorithme à résoudre le problème pour une paire d'agents génétiquement identiques.

A nouveau, les agents ne reçoivent aucune information ni des caractéristiques de la session en cours ni même du moment où commence une nouvelle session. Les agents doivent s'adapter à leur environnement (à l'autre agent en l'occurrence). Pour cette tâche, la population d'agents est initialisée aléatoirement. L'étape de sélection et les probabilités d'applications des opérateurs sont identiques à la première tâche. Pendant 1 000 itérations, un agent A rencontre un autre agent B.

3.1.3. *Le contrôle mutuel du destin avec sélection comportementale*

Cette tâche est strictement identique à la procédure précédente mis à part en ce qui concerne la population initiale. Elle n'est pas aléatoire mais constituée d'agents capables de réussir une procédure de test. Cette procédure de test consiste en un test d'apprentissage très simple. Elle consiste à renforcer certains comportements. S'ils apparaissent plus fréquemment ensuite, on considère que l'agent a réussi le test.

Dans la pratique, on observe le comportement « spontané » (1 000 cycles activation/apprentissage) du réseau et l'on choisit de renforcer l'un des comportements par l'activation de l'un des neurones de la couche d'entrée. Si, suite à ce renforcement, le comportement est répété davantage qu'un comportement non renforcé, on considère que l'épreuve est réussie. On ne retient que les agents qui réussissent 10 épreuves successives. Au cours de chaque épreuve, on renforce alternativement le comportement B1 ou B2. De cette façon, on évite de retenir des agents qui n'émettent que B1.

Au total, 813 agents aléatoires ont été nécessaires pour constituer une population initiale de 10 agents qui réussissent le test. Une fois constituée, la population effectue la tâche du contrôle mutuel du destin.

3.2. ***Résultats des simulations***

Les agents et les procédures décrites ont été implantés en Java^{tm} pour réaliser les simulations. Cette section présente les résultats de ces simulations, tâche par tâche.

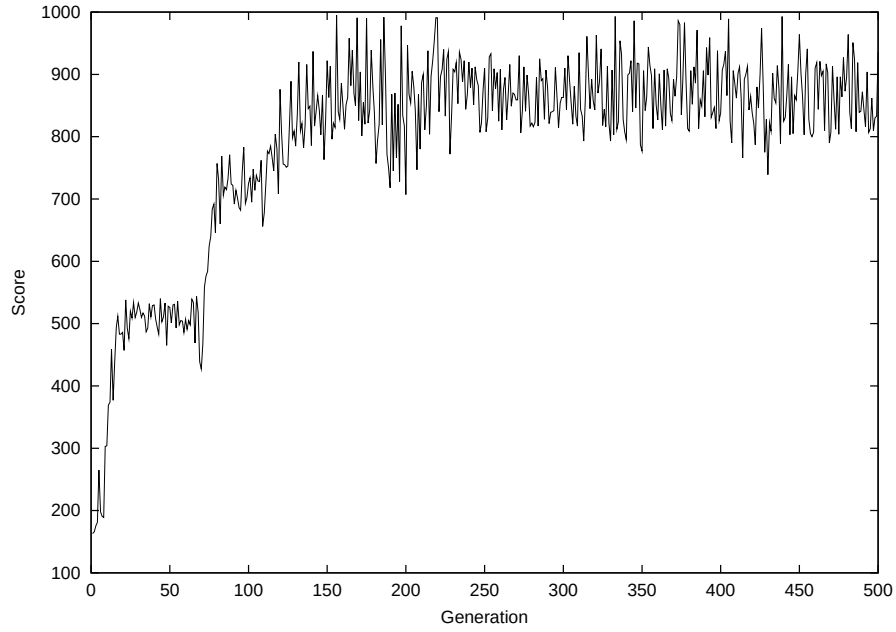

Figure 4. *Performance moyenne en fonction du temps des agents confrontés à la tâche de discrimination. La performances enregistrée est le nombre de renforcements reçus. Le maximum est 1 000*

3.2.1. *Tâche de discrimination*

La figure 4 représente la performance moyenne de la population d'agents dans la tâche de discrimination au cours du temps. On constate qu'après une augmentation rapide de la performance des agents, celle-ci se stabilise sur un palier (proche de 500) après quoi elle réaugmente pour atteindre un nouveau palier plus élevé (environ 850). Le premier palier correspond à une population dans laquelle les agents sont capables de recevoir le renforcement une fois sur deux : la performance est de 500 alors que le maximum vaut 1 000 (puisque 1 000 stimuli sont présentés à chaque agent, voir paragraphe 3.1.1). Ainsi, les agents de cette population ont appris une association sur deux. Ensuite, dans cette population, la capacité à discriminer apparaît rapidement. Après 200 générations, environ 90 % des agents sont capables de réaliser la tâche de discrimination. Chaque agent est alors capable de réagir à chacune des associations et de s'adapter aux changements.

La simulation montre que l'évolution génétique peut retenir la capacité à discriminer, c'est-à-dire à apprendre à répondre en fonction des conséquences de ces réponses. Il faut insister sur le fait que l'environnement est dynamique : les renforçateurs ne sont pas reçus après l'émission du même comportement ; rien dans l'environnement ne permet à l'agent de savoir dans quelle condition il se trouve à un moment donné.

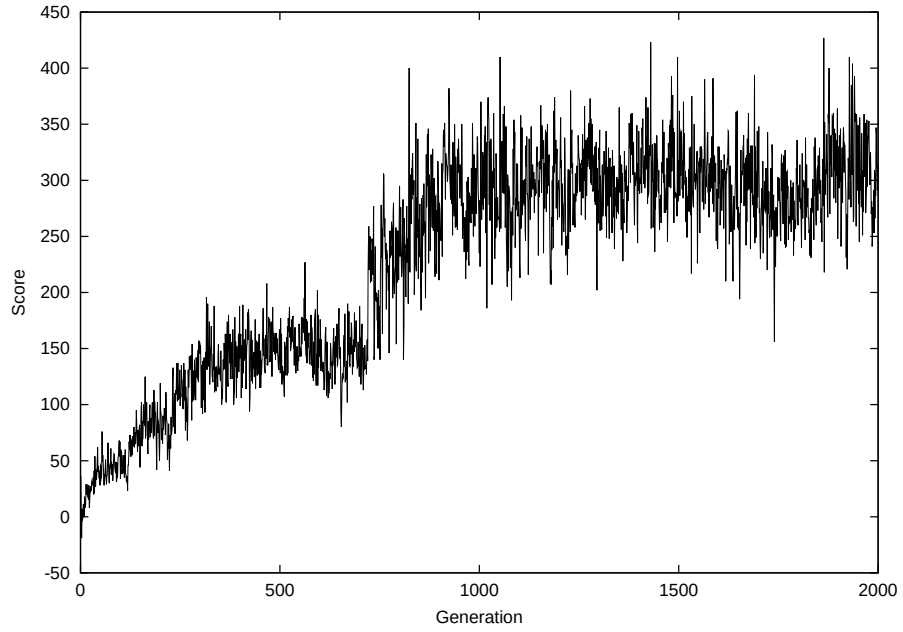

Figure 5. *Performance moyenne en fonction du temps des agents confrontés à la tâche MFC. La performance mesurée est le nombre de renforcements reçus. Le maximum est 1000. La population initiale est constituée d'agents choisis aléatoirement*

La capacité à émettre un comportement qui s'est avéré profitable par la suite constitue le principe même de la loi de l'effet et par conséquent du principe de sélection des comportements par leurs conséquences. Donc cette simulation suggère que ce principe peut être le résultat de la sélection naturelle. Sur cette base, la simulation suivante montre que la capacité à apprendre procure un avantage important.

3.2.2. *Contrôle mutuel du destin*

La figure 5 présente l'évolution de la performance moyenne des agents dans la situation de contrôle mutuel du destin. De façon nette, l'évolution génétique aboutit à augmenter la capacité d'un agent à contrôler le comportement d'un autre agent. Cette évolution se fait en plusieurs phases. Des augmentations soudaines de la performance sont observées, parfois séparées par de périodes de stabilité. Il faut cependant remarquer qu'après 2 000 générations, la performance des agents reste modeste : ils reçoivent seulement 35 % des renforçateurs qu'ils pourraient obtenir.

3.2.3. *Contrôle mutuel du destin avec sélection comportementale*

Enfin, la population initiale des agents qui sont confrontés à la situation du contrôle mutuel du destin est composée d'agents qui ont réussi la procédure de test décrite

plus haut. Dans ce cas, l'évolution de la performance est très différente. La figure 6 montre cette différence entre l'évolution de la population constituée d'agent aléatoire (coubre A) et l'évolution de la population lorsque la population initiale réussit le test (courbe B). Au début, la performance des deux populations est proche. Cependant, après quelques générations, la population B se montre nettement meilleure que la population A. Après 200 générations, la population B obtient 85-90 % des renforçateurs disponibles. L'écart à la valeur optimale est dû, notamment, au fait qu'au début de chaque session, les agents doivent réadapter leur comportement. Une autre part de l'écart à l'optimum est liée au fait que l'agent maintient en permanence une part d'exploration dans son comportement. Dans la mesure où aucune information n'est donnée sur le « bon » comportement, la stratégie adaptative consiste à explorer de temps à autre des solutions non optimales. Ces deux facteurs résultant du processus adaptatif expliquent l'écart de 10 % à l'optimum. Constatons qu'on observe cette même sous-optimalité du comportement des organismes vivants ; de même, l'exploration est un élement important des algorithmes d'apprentissage par renforcement [SUT 98] en complément de l'exploitation des solutions déjà trouvées.

4. Conclusion et discussion

Dans cet article, nous avons présenté un modèle et ses simulations ayant pour objectif de montrer que la capacité d'apprendre peut résulter de l'évolution génétique. Cette capacité apparaît par sélection, *via* l'effet Baldwin. L'apprentissage implique qu'une certaine structure des agents est capable d'apprendre les associations entre stimulus et comportement à émettre. L'apprentissage ne signifie pas pour autant l'acquisition d'un réflexe de type « stimulus-réponse » : l'environnement est dynamique et les agents doivent donc être capables de modifier leur comportement à tout moment au cours de leur « vie ». Ceci correspond typiquement à un conditionnement opérant ou instrumental (apprentissage par renforcement). Ce travail doit donc être considéré comme le prolongement des recherches concernant les algorithmes évolutionnaires : la sélection naturelle peut produire des individus de plus en plus adaptés au cours des générations à la fois dans des environnement statiques et dynamiques. En tant que tel, ce point n'est cependant pas original ; l'originalité tient à ce que nous avons essayé de minimiser les capacités des agents mis en jeu ; on souhaite ainsi éviter de mettre la solution dans l'énoncé du problème en utilisant des réseaux de neurones supervisés ou même des algorithmes de renforcement. Au contraire, on souhaite que les capacités exhibées par un algorithme de renforcement émergent de l'évolution d'agents non supervisés. Ainsi, les agents réussissant la tâche de discrimination sont-ils des algorithmes de renforcement élémentaires. Dès lors, la synthèse d'agents correspondant à des algorithmes de renforcement plus sophistiqués n'est vraisemblablement qu'une question de complexification de la simulation (pas du modèle) et de temps d'exécution.

Insistons sur le fait qu'en aucun cas nous n'avons cherché ici à obtenir un résultat optimal, quel que soit le sens que l'on veuille donner à cet adjectif : taille du réseau

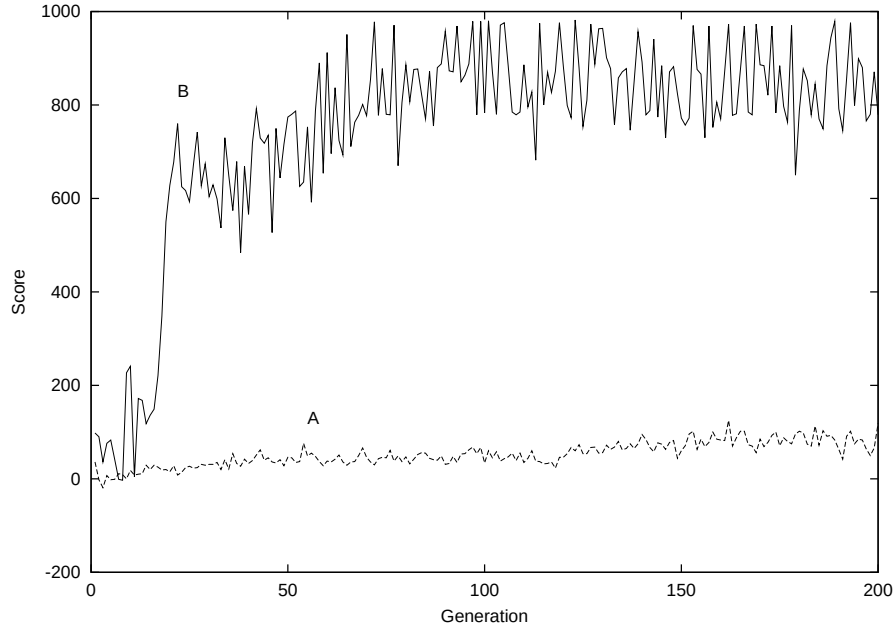

Figure 6. *Evolution de la performance au cours du temps des deux populations placées en situation de contrôle mutuel du destin : la population A est composée initialement d'agents déterminés aléatoirement (en fait, la même population que celle de la figure 5 dont on présente ici uniquement les 200 premières générations) ; B est une population constituée initialement d'agents qui réussissent un test très simple d'apprentissage. Une réussite, même faible, à ce test avantage nettement cette population qui obtient de bonnes performances après 200 générations dans la tâche de contrôle mutuel du destin. Bien que l'environnement se modifie brutalement à la fin de chaque session, la performance est de 85 à 90 %*

de neurones utilisé, sa structure, le temps d'apparition d'un agent réalisant la tâche, les paramètres des algorithmes (taille de la population, taux d'application des opérateurs…), opérateurs génétiques utilisés…Ces points, qui pourraient légitimement être étudiés en tant que tels, sont totalement étrangers à la démarche que nous avons suivie ici. Il faut donc relier ce travail à celui de différents auteurs étudiant les processus d'apprentissage et l'interaction entre sélection naturelle, apprentissage, voire culture. Ce faisant, nous synthétisons une architecture de renforcement qui peut être utilisée en place d'autres algorithmes de renforcement. On peut ainsi envisager de comparer les réseaux obtenus avec un algorithme de type Q-Learning. Par ailleurs, une étude à faire consisterait, dans un premier temps, à étudier les réseaux obtenus dans nos simulations pour en détecter certaines propriétés caractéristiques invariantes qui en feraient des architectures de renforcement et, dans un second temps, à essayer de synthétiser des réseaux de taille arbitraire vérifiant ces propriétés, réseaux qui seraient alors ca-

pables de traiter un nombre quelconque de stimuli en entrée et d'émettre un nombre arbitraire de comportements en sortie. La découverte d'une structure invariante dans ces réseaux permettraient d'éviter le recours à une phase d'évolution génétique pour synthétiser ces réseaux. Cependant, la détection de caractéristiques invariantes est un point qui semble *a priori* difficile.

Un point important est que nous envisageons l'adaptation dans un environnement non stationnaire. Le fait que l'agent soit confronté à un environnement dynamique est simplement simulé en le faisant interagir avec un autre agent. Chaque agent rencontre un clone de lui-même : un individu qui a le même génome ; cependant, bien qu'ayant le même génome, les deux individus n'ont pas forcément le même comportement, tous comme deux jumeaux. Ce choix est dicté par la tâche, très spécifique, que nous utilisons et dans laquelle le comportement de l'agent n'affecte pas son propre gain. Ainsi, nous évitons que des agents bons dans cette tâche ne soit éliminés par d'autres plus mauvais. Il est probable que cette manière de faire ne soit pas nécessaire mais qu'elle accélère seulement l'évolution des réponses.

L'un des problèmes que nous avons évoqués est celui de la plausibilité biologique des modèles que nous mettons en œuvre. Dans la mesure où le vivant est utilisé comme modèle d'adaptation à l'environnement, on recherche souvent à développer des architectures et des mécanismes proches de ceux décrits biologie. Cela dit, nous recherchons une plausibilité fonctionnelle car notre objectif est de réaliser des modèles qui présentent des comportements adaptatifs. La reproduction pure et simple, dans tous ses détails, d'une structure biologique n'a pas d'intérêt. En revanche, s'il est démontré qu'une certaine structure entraîne une modification du comportement (meilleure exploration, adaptation à des environnements dynamiques…), il est nécessaire de la reproduire.

On constate qu'une fois acquise la capacité à sélectionner les comportements en fonction de leurs conséquences, l'interaction entre agents devient possible [DEL 00b] et que des comportements complexes apparaissent [PRE 01, DEL 01b]. Une autre particularité de ce travail est de mettre en avant les interactions entre deux agents adaptatifs et pas seulement l'adaptation d'un seul agent dans un environnement statique ou dynamique. Ceci est un point important pour obtenir des modèles et des simulations réalistes des phénomènes vivants.

5. Bibliographie

[ACK 92] ACKLEY D., LITTMAN M., « Interactions between learning and evolution », *in [LAN 92]*, 1992, p. 487–509.

[BAL 96] BALDWIN J., « A new factor in evolution », *The american naturalist*, vol. 30, 1896, reprinted in [MIT 96a, pp. 59-80].

[BEL 90] BELEW R. K., « Evolution, Learning and Culture : Computational Metaphors for Adaptive Algorithms », *Complex Systems*, vol. 4, 1990, p. 11–49.

[BON 99] BONABEAU E., DORIGO M., THÉRAULAZ G., *Swarm Intelligence : From natural to Artificial Systems*, Oxford University Press, 1999.

[CHA 90] CHALMERS D. J., « The Evolution of Learning : An Experiment in Genetic Connectionism », TOURETZKY D., ELMAN J., SEJNOWSKI T., HINTON G., Eds., *Proc. of the 1990 Connectionist Models Summer School*, Morgan Kaufmann, San Mateo, CA, USA, 1990, also available as report 48 of the CRCC, Indiana University, Bloomington, IN 47405, USA.

[CHA 99] CHANCE P., « Thorndike's puzzle boxes and the origins of the experimental analysis of behavior », *Journal of the Experimental Analysis of Behavior*, vol. 72, n° 3, 1999, p. 433-440.

[DEL 00a] DELEPOULLE S., « Coopération entre agents adaptatifs ; étude de la sélection des comportements sociaux, expérimentations et simulations », PhD thesis, Université de Lille 3, URECA, Villeneuve d'Ascq, octobre 2000, Thèse de doctorat de Psychologie.

[DEL 00b] DELEPOULLE S., PREUX P., DARCHEVILLE J.-C., « Dynamics of Temporal Organization of Behaviors in Interaction Situation », 2000, (submitted).

[DEL 00c] DELEPOULLE S., PREUX P., DARCHEVILLE J.-C., « Evolution of cooperation within a behavior-based perspective : confronting nature and animats », *Artificial Evolution'99*, vol. 1829 de *Lecture Notes in Computer Science*, Springer-Verlag, 2000, p. 204–216.

[DEL 01a] DELEPOULLE S., PREUX P., DARCHEVILLE J.-C., « Dynamique de l'interaction », CHAIB-DRA B., ENJALBERT P., Eds., *Proc. Modèles Formels de l'Interaction, Toulouse*, 2001, p. 141–150.

[DEL 01b] DELEPOULLE S., PREUX P., DARCHEVILLE J.-C., « Selection of behavior in social situations — Application to the development of coordinated movements », *Applications of Evolutionary Computing*, vol. 2037 de *Lecture Notes in Computer Science*, Springer-Verlag, avril 2001, p. 384–393.

[FLO 93] FLOREANO D., « Emergence of Nest-Based Foraging strategies in Ecosystems of Neural Networks », MEYER J., ROITBLATT H., WILSON S., Eds., *Proc. SAB 2*, MIT Press, 1993, p. 410–416.

[FLO 96] FLOREANO D., MONDADA F., « Evolution of plastic neurocontrollers for situated agents », MAES P., MATARIC M., MEYER J., POLLACK J., ROITBLATT H., WILSON S., Eds., *Proc. SAB 4*, MIT Press, 1996, p. 402–410.

[FLO 99] FLOREANO D., NOLFI S., « Learning and evolution », *Autonomous robots*, vol. 7, n° 1, 1999, p. 89–113.

[FOG 66] FOGEL L. J., OWENS A. J., WALSH M. J., *Artificial Intelligence Through Simulated Adaptation*, Wiley, New York, 1966.

[FOR 91] FORREST S., « Emergent computation : self-organizing, collective, and cooperative phenomena in natural and artificial computing networks », FORREST S., Ed., *Emergent Computation*, A Bradford Book, p. 1–11, MIT Press, 1991.

[GRU 92] GRUAU F., « Genetic systems of boolean neural networks with a cell rewriting developmental process », WHITLEY D., SCHAFFER J., Eds., *Combinaiton of Genetic Algorithms and Neural Networks*, IEEE Computer society press, 1992.

[HIN 87] HINTON G., NOWLAN S., « How Learning Can Guide Evolution », *Complex Systems*, vol. 1, 1987, p. 495–502, also reproduced in [MIT 96a], chapter 25, pp. 447-454.

[HOL 61] HOLLAND J. H., « Outline of a Logical Theory of Adaptative Systems », *Journal of the ACM*, vol. 7, 1961, p. 297–316.

[HOL 75] HOLLAND J. H., *Adaptation in Natural and Artificial Systems*, Michigan Press University, Ann Arbor, MI, 1975.

[KOD 98] KODJOBACHIAN J., MEYER J., « Evolution and development of neural controllers for locomotion, gradient-following, and obstacle avoidance in artificial insects », *IEEE Transactions in Neural Networks*, vol. 9, 1998, p. 796–812.

[LAN 92] LANGTON C., TAYLOR C., FARMER J. D., RASMUSSEN S., Eds., *Artificial Life II*, SFI Studies in the Sciences of Complexity, Addison-Wesley, 1992.

[LIT 96] LITTMAN M., « Simulations combining evolution and learning », *in [MIT 96a]*, p. 465–477, 1996.

[MET 53] METROPOLIS N., ROSENBLUTH A., ROSENBLUTH M., TELLER A., « Equations of state calculations by fast computing machines », *Journal of Chemical Physics*, vol. 21, 1953, p. 1087–1092.

[MIG 96] MIGLINO O., NOLFI S., PARISI D., « Discontinuity in evolution : how different levels of organizaiton imply pre-adaptation », *in [MIT 96a]*, 1996.

[MIT 96a] MITCHELL M., BELEW R., Eds., *Adaptive Individuals In Evolving Population Models*, SFI Studies in the Sciences of Complexity, Addison-Wesley, 1996.

[MIT 96b] MITCHELL M., *An Introduction to Genetic Algorithms*, MIT Press, A Bradford Book, 1996.

[MOR 96] MORGAN C. L., « On modification and variation », *Science*, vol. 4, 1896, p. 733–740.

[OSB 96] OSBORN H., « Ontogenetic and phylogenetic variation », *Science*, vol. 4, 1896, p. 786–789.

[PAR 96] PARISI D., NOLFI S., « The influence of learning on evolution », *in [MIT 96a]*, p. 419–428, 1996.

[PIT 47] PITTS W., MCCULLOCH W., « How we know universals : the perception of auditory and visual forms », *Bulletin of Mathematical Biophysics*, vol. 9, 1947, p. 127–147.

[PRE 01] PREUX P., DELEPOULLE S., DARCHEVILLE J.-C., « Selection of behaviors by their consequences in the human baby, software agents, and robots », *Proc. Computational Biology, Genome Information Systems and Technology*, mars 2001.

[REC 73] RECHENBERG I., *Evolutionsstrategie : Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*, Frommann-Holzboog Verlag, Stuttgart, 1973.

[SID 56] SIDOWSKI J., WYCKOFF B., TABORY L., « The influence of reinforcement and punishment in a minimal social situation », *Journal of Abnormal Social Psychology*, vol. 52, 1956, p. 115–119.

[SKI 38] SKINNER B., *The behavior of organisms*, Appleton-Century Crofts, 1938.

[SKI 81] SKINNER B., « Selection by consequences », *Science*, vol. 213, 1981, p. 501–514.

[STA 00] STADDON J., *The new behaviorism — Mind, Mechanism, and Society*, Psychology Press, 2000.

[SUT 98] SUTTON R., BARTO A., *Reinforcement learning : an introduction*, MIT Press, 1998.

[THO 98] THORNDIKE E., « Animal Intelligence : An experimental study of the associative process in animals », *Psychology Monographs*, vol. 2, 1898.

[THO 11] THORNDIKE E., *Animal Intelligence : Experimental Studies*, Mac Millan, 1911.

[URZ 00] URZELAI J., « Evolutionary Adaptive Robots : artificial evolution of adaptation mechanisms for autonomous systems », PhD thesis, EPFL, Lausanne, Suisse, 2000.

[WAD 53] WADDINGTON C., « Genetic assimilation for acquired character », *Evolution*, vol. 7, 1953, p. 118–126.

[WAD 56] WADDINGTON C., « Genetic assimilation of the *bithorax* phenotype », *Evolution*, vol. 10, 1956, p. 1–13.

[WIL 75] WILSON E., *Sociobiology*, Belknap, Harvard University Press, 1975.