Adaptation réflexive : Contrôleur de comportement et structure évolutive

Elpida S. Tzafestas
Institute of Communication and Computer Systems
National Technical University of Athens
Heroon Polytechneiou 9
Zographou 15773, Greece.
brensham@softlab.ece.ntua.gr
http://www.softlab.ece.ntua.gr/~brensham

Résumé

Dans cet article, nous présentons le concept de l'adaptation réflexive comme déduit depuis trois modèles d'agents autonomes conçus spécifiquement pour des domaines différents de comportement : un agent explorateur adaptatif, une fourmis qui communique avec ses soeurs à l'aide de pheromones, un agent coopératif adaptatif. Ainsi nous comparons ces trois modèles en ce qui concerne leur mode d'adaptation en dehors de leur contexte détaillé d'opération. Nous constatons que dans les trois cas, il est question de variables "cognitives" constamment mises-à-jour et d'un mécanisme d'adaptation de deux niveaux, le premier agissant sur la variable cognitive et le deuxième sur le système d'adaptation de premier niveau. Le jeu entre les dynamiques différentes des deux niveaux d'adaptation est responsable de la performance finale de l'agent dans son environnement. Mais l'adaptation réflexive n'est pas simplement le contrôleur du comportement; elle peut servir également de structure évolutive. Plus précisément, nous pouvons imaginer une structure "paramétrée" qui sera adaptée de génération à génération de façon évolutive. Nous esquissons cette structure et nous discutons sa signification : d'une part, elle assure la convergence du comportement quant à ses performances, et d'autre part elle reste invariante à travers les problèmes posés et les applications.

1 Introduction : Adaptation

Dans cet article, nous passons en revue et nous comparons trois modèles d'agents autonomes qui résoudent trois problèmes typiques, l'agent explorateur (Tzafestas 1995, soumis 1), l'agent-fourmis (Tzafestas 1998, soumis 2) et l'agent tit-for-tat coopératif (Tzafestas 2000). Dans tous les cas, et malgré les différences apparentes entre eux, nous avons réussi à définir une mesure interne à l'agent et dépendante du problème qui représente l'état d'avancement de la résolution et un mécanisme d'autorégulation qui possède deux parties couplées. Le but de l'agent est de réguler la valeur de sa variable interne entre des limites, soit pour la ramener à une valeur de référence, telle que 0, soit pour l'empêcher de toucher les extrêmes. La régulation est positive, de manière que la variable de l'agent suive la tendance de son environnement qu'elle essaye de représenter, même si la valeur de cette variable ne correspond jamais à la réalité, mais elle assure une distance représentationnelle. À un deuxième niveau méta, un autre mécanisme de régulation "regarde" le premier niveau et régule ses taux d'adaptation de manière négative. La

puissance de l'agent réside dans l'existence de ce mécanisme de régulation de deuxième niveau qui est dépendant du problème et qui a conduit à une amélioration de la performance de l'agent par rapport au modèle de base d'un seul niveau. La conception générale de l'ensemble de ce système physiologique peut être résumée ainsi :

Si (le monde diverge de la représentation de l'agent)
alors (dans le futur) adapter afin de venir plus près au monde,
sinon(dans le futur) adapter afin d'amplifier cette différence.

L'adaptation devient ainsi réflexive dans le sens que le deuxième niveau d'adaptation regarde et agit sur le premier, autrement dit le système complet d'adaptation comme vu par l'extérieur se regarde et s'adapte de manière intrinsèque.

2 Étude de cas I : L'agent explorateur

2.1. Le problème

Un problème typique de robotique comportementale est celui de ***l'exploration*** : un ensemble d'agents (robots) débarque sur une planète avec la mission d'explorer sa surface pour des échantillons de minerai ayant certaines propriétés. Les robots arrivent dans un vaisseau spatial qui sert de base planétaire tout au long de la mission. La mission est accomplie quand toute la surface dans un certain rayon de la base est explorée, c'est-à-dire quand les agents ont "balayé" toute la surface en question et ont ramassé toutes les instances d'objets d'intérêt (voir par exemple Beckers et al. 1994). Les agents sont supposés de rentrer à la base lorsque leur mission est finie.

Ce problème d'exploration a été traditionnellement abordé du point de vue "fonctionnel" : "Comment un ou plusieurs agents balayent un espace délimité pour épuiser les sources d'intérêt ?". La réponse à cette question est un système de contrôle, une architecture, qui permet à l'agent de naviguer, percevoir, détecter du minerai etc., afin de balayer tout l'espace en question. Une solution comme celles rencontrées dans la bibliographie (par exemple Mataric 1992), avec un composant aléatoire et même sans apprentissage ou raisonnement spatial, assure statistiquement la couverture du champ d'intérêt et l'épuisement des sources de minerai.

Mais d'un point de vue plus "cognitif", cette fonctionnalité seule ne répond pas à la question essentielle : "***Comment les agents savent-ils qu'ils ont balayé tout l'espace, ou qu'ils ont accompli leur mission ?***". Pour répondre à cette question, il faut reformuler la description de la tâche de balayage, de manière à y inclure une expression, analytique ou autre, qui représente le critère de terminaison, c'est-à-dire l'épuisement des sources de minerai. Il suffit alors de définir une variable environnementale, la densité des sources de minerai (dénotée p_m par la suite), qui caractérise l'état du monde à un instant donné. Le but de l'agent explorateur-balayeur devient donc de ramener la valeur de cette variable à 0. Nous verrons qu'un agent ayant une représentation de cette variable constitue une solution simple à ce problème de description.

Troisièmement, nous cherchons à étudier l'opérationalité du système, c'est-à-dire la relation entre l'architecture interne des agents et leurs performances, dans le but de trouver une architecture qui "optimise" ces performances-là. Le critère d'opérationalité qui s'applique à la tâche de balayage est, bien évidemment, la durée de la mission : les agents sont plus performants s'ils se rendent compte de la terminaison de leur mission plus rapidement.

Dans les simulations effectuées, le monde sous exploration est défini comme un carré autour de la base centrale : la taille du monde est alors la longueur du carré (par défaut, les résultats reportés par le suite ont été pris dans un monde 25x25). Le système de contrôle de base de l'agent ainsi que les détails de simulation sont donnés dans (Tzafestas 1995, soumis 1), pour les deux cas d'un seul agents et d'agents multiples.

2.2. La solution : Reformulation du problème comme un problème d'adaptation

Nous abordons maintenant la deuxième question : "Comment les agents savent-ils qu'ils ont balayé tout l'espace pour rentrer définitivement à la base ?". Ils ont besoin d'un moyen de détection du degré de complétion de la tâche ou d'un critère de terminaison (balayage complété). Le seul paramètre de la tâche qui peut être utile pour le développement d'un critère de terminaison est la densité des sources dans le monde $p_m(t)$. Si l'agent connaissait d'avance sa valeur initiale $p_m(0)$, on pourrait définir comme critère de terminaison une formule du type *{$p_m(0)$ * sqr(r) échantillons ont été ramassés}* (où r est la longueur du côté du carré, ici 25). Cependant, ce critère n'est pas sûr, parce que, si un échantillon n'est pas détecté, l'agent ne terminera jamais (mais on pourrait laisser tomber un ou deux échantillons qu'on n'a pas trouvés).

La solution très simple à ce problème est d'estimer continûment la valeur de $p_m(t)$ et, étant donné qu'elle tombe à 0 comme effet de bord de l'activité de l'agent, de prendre comme critère de terminaison $p_m(t)=0$. L'estimation de la valeur de $p_m(t)$ nécessite alors une variable représentationnelle locale à l'agent ($p_a(t)$) et peut se faire par l'intermédiaire d'une formule simple d'adaptation proportionnelle :

Variable représentationnelle : $p_a(t)$

Adaptation proportionnelle :
fenêtre d'observation w, taux r

$p_a(t) = p_a(t\text{-}w) + diff * r$

$diff = p_{calc} - p_a(t\text{-}w)$

p_{calc} = *nombre des échantillons ramassés / nombre des pas effectués (pendant la fenêtre de l'adaptation)*

Critère de terminaison :
$p_a(t) < e_p$
où e_p un petit seuil (ici, e_p=0.001)

Le p_{calc} exprime l'estimation de l'agent lors de sa fenêtre d'observation et la loi proportionnelle assure que la mise-à-jour de l'estimation de l'agent ne se fait pas trop rapidement. Ce système de représentation et d'adaptation présente l'avantage de robustesse face aux perturbations/manipulations du type réinitialisation de la variable $p_m(t)$ au cours du balayage. Dans la figure 1 est illustrée la co-évolution des deux variables $p_m(t)$ et $p_a(t)$ dans le temps. Comme on peut le voir sur la figure, ***la variable représentationnelle permet à l'agent de résoudre son problème de terminaison dans tous les cas sans jamais prendre la valeur réelle qu'elle représente*** (sauf un point de croisement). Les deux variables tombent progressivement à 0 sans jamais prendre la même valeur — on pourrait dire que celle de $p_a(t)$ "suit" celle de $p_m(t)$. En effet, la montée rapide de $p_a(t)$ au début du balayage est la conséquence de l'utilisation d'un capteur de détection d'échantillons à distance qui fait orienter l'agent vers les

sources de minerai en minimisant son comportement erratique de façon que la plupart des places visitées soient des places contenant des échantillons. La valeur de $p_a(t)$ baisse ensuite puisque celle de $p_m(t)$ baisse comme effet de bord de l'activité de l'agent qui trouve de moins en moins des échantillons.

2.3. Adaptation à deux niveaux

Nous avons ensuite voulu étudier la relation — s'il en existe une — entre les paramètres *w* et *r* du système d'adaptation et la valeur $p_m(0)$. Le système a été alors simulé pour des différentes valeurs de *w* et *r* et dans des différentes densités initiales de monde. La figure 1 donne les résultats de ces simulations pour trois ensembles des paramètres d'adaptation (adaptation rapide, moyenne ou lente).

L'adaptation rapide est plus opérationnelle que l'adaptation moyenne qui est à son tour plus opérationnelle que l'adaptation lente (toujours selon le critère de durée). Cependant, plus l'adaptation est rapide, plus elle démontre des fluctuations, et plus l'adaptation est lente, plus elle démontre des retards. Qui plus est, le même paramétrage donne des résultats différents dans les différentes densités de monde : la différence des résultats se voit dans la forme des courbes (pour plus de résultats voir Tzafestas 1995, soumis 1). Plus particulièrement, la réponse de l'agent aux différentes perturbations (la forme de la courbe d'évolution de $p_a(t)$) diffère selon la condition de limite $p_m(0)$: pour le même paramétrage d'adaptation, l'agent finit plus ou moins vite sa mission selon la valeur de $p_m(0)$, c'est-à-dire l'intervalle entre le moment du ramassage du dernier échantillon et le retour définitif de l'agent à la base est d'une durée très variable. Il semble donc que, pour assurer l'opérationalité de l'agent dans les différents mondes, il faut trouver un moyen de combiner les avantages de l'adaptation rapide en termes d'opérationalité avec les avantages de l'adaptation lente en termes de régularité de courbe.

Plus précisément, il faut une adaptation rapide à la fin (pour terminer rapidement), mais lente lors du ramassage (pour éviter les fluctuations). Il s'agit alors de trouver un moyen de se stabiliser au bon paramétrage *en ligne*. Autrement dit, ***il faut un système de méta-adaptation***.

Méta-adaptation :
Si $|diff|$ $(= |p_{calc}-p_a(t-w)|) \leq f_p$,
alors adaptation plus rapide
$r => r_{max}$, $w => w_{min}$
sinon adaptation plus lente
$r => r_{min}$, $w => w_{max}$
$r = r + r_r * (r_{max}-r)$, $w = w + r_w * (w_{min}-w)$

La méta-adaptation doit affecter les paramètres *w* et *r* de façon que l'adaptation devienne plus rapide quand le p_{calc} est suffisamment près de $p_a(t)$ et plus lente quand il est plus loin. Cette loi de méta-adaptation signifie que le monde paraît plus fiable quand il ne diffère pas trop de l'idée que l'agent en a, sinon il est pris moins au sérieux.

La figure 2 donne les résultats de l'application du système de méta-adaptation pour les trois densités du monde ; comme on peut facilement voir sur la figure, la réponse de l'agent (la forme de la courbe) est la même pour les trois densités exemplaires, autrement dit le résidu de

la durée de la mission après le ramassage du dernier échantillon est approximativement le même dans les trois cas.

Nous avons montré dans (Tzafestas 1995, soumis 1) que l'opérationalité de l'agent qui possède ce système de méta-régulation ne dépend pas qualitativement des valeurs des paramètres w_{min}, w_{max}, r_{min}, r_{max}, r_w et r_r. De plus, la condition initiale $p_a(0)$ n'a aucune importance non plus.

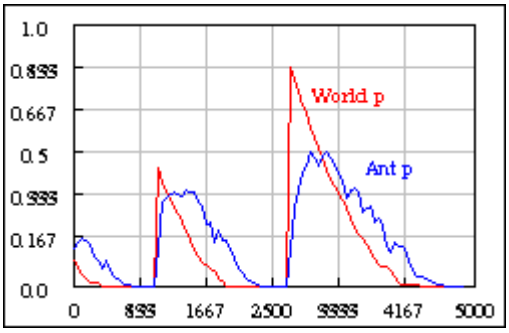

Figure 1. Performance de l'agent adaptatif de base pour trois densités initiales de monde, densité faible ($p_m(0)$=0.1), moyenne ($p_m(0)$=0.1) et élévée ($p_m(0)$=0.9) ($p_a(0)$=0.15). Moments de terminaison 1ère partie = 1019, 2ème partie = 2659, 3ème partie = 4897). (f_p=0.1, w=30, r=0.2, adaptation moyenne)

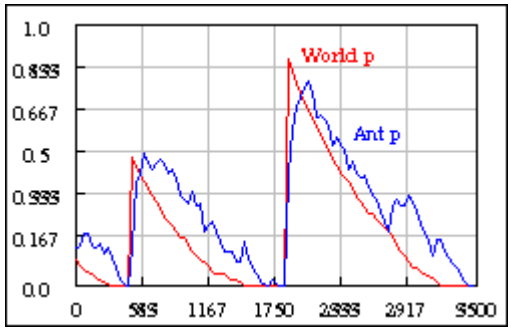

Figure 2. Performance de l'agent adaptatif avec du méta pour trois densités initiales de monde, densité faible ($p_m(0)$=0.1), moyenne ($p_m(0)$=0.1) et élévée ($p_m(0)$=0.9) ($p_a(0)$=0.15). Moments de terminaison 1ère partie = 489, 2ème partie = 1832, 3ème partie = 3497). (f_p=0.1, w_{min}=15, w_{max}=40, r_{min}=0.15, r_{max}=0.3, r_r=r_w=0.2)

3 Étude de cas II : La fourmis et ses phéromones

3.1. Le problème

La situation naturelle rencontrée dans les sociétés d'insectes et souvent modélisée est une variante du problème précédent où il existe peu de sources de taille importante distribuées dans l'environnement. La solution dans ce cas consiste à permettre au robot-fourmis de déposer de pheromones lorsqu'il se trouve chargé et de retour à sa base, c'est-à-dire à son nid. Un autre agent ou lui-même peut suivre ces traces pour arriver à la source d'intérêt rapidement. Une possibilité supplémentaire est de considérer que la trace de pheromones déposées par terre évapore lentement (Deneubourg et al. 1990, Solé et al. 2000).

La première motivation de notre approche est que le modèle ainsi défini ne doit pas être stable pour cause de dépendre d'une quantité infinie de pheromone si la nourriture est régulièrement renouvelée. Ce problème est étudié en détail dans (Tzafestas 1998, soumis 2).

Le modèle comportemental de base (Steels 1990, Drogoul et Ferber 1992) est le suivant :

Si (de retour au nid, c'est-à-dire chargé)
Si (sur le nid) décharger
Sinon {aller vers le nid, déposer 2 unités de pheromone}
Sinon Si (sur une source) charger un morceau
Sinon Si (des pheromones ou un stimulus perçu)
{suivre le stimulus, ramasser 1 unité de pheromone}
Sinon aller au hasard

Le modèle de base dépose 2 unités de pheromone quand il retourne à sa base et ramasse 1 unité quand il suit la pheromone déjà sur le terrain ou un autre stimulus. Dans les simulations

effectuées, le monde sous exploration est défini comme un carré autour de la base centrale : la taille du monde est alors la longueur du carré (par défaut, les résultats reportés par le suite ont été pris dans un monde 20x20). La source de nourriture se trouve dans le coin de droite en bas, le nid se trouve dans le coin de gauche en haut et la population des robots-fourmis consiste de 5 robots. Les robots peuvent percevoir la nourriture ou la pheromone depuis une distance de 3 unités d'espace. Nous avons simulé ce système avec le comportement de base d'agent-fourmis et nous avons mésuré les quantités de pheromone dans le monde et dans chaque agent. Comme prévu, les quantités de pheromone des agents tombent sous 0, tandis que la quantité de pheromone dans le monde peut croître sans limite. Les valeurs de ces quantités dépendent des autres paramètres du problème (distance source-nid, nombre de robots, taille de source etc.) qui définissent le nombre nécessaire des voyages aller-retour source-nid qui sont nécessaires pour épuiser la source. Les mêmes résultats sont obtenus quand l'agent-fourmis ne ramasse pas de pheromone, mais elle évapore naturellement, ou si la pheromone possédée par chaque agent regénère naturellement aussi. Un deuxième problème émerge si l'on essaye de résoudre le premier (valeurs négatives) de façon non naturelle, en dotant depuis le départ les agents d'une quantité suffisante de pheromone, c'est-à-dire très grande. Ceci conduit à un système où les agents continuent à être attirés par la grande quantité de pheromone à une source de nourriture déjà épuisée depuis longtemps.

3.2. La solution : Reformulation du problème comme un problème d'adaptation

Les observations précédentes nous amèment à une nouvelle formulation du problème de marquage par pheromones : ***Nous cherchons un mécanisme de régulation de pheromone qui permet au chemins de pheromone d'être construits et renforcés rapidement autant que la source est présente et disparaître aussitôt après son épuisement***.

Un mécanisme simple qui permet la régulation de pheromone entre des bornes utilise deux valeurs $pheromones_{min}$ et $pheromones_{max}$ et la "loi d'effet" :

Pour déposition $pheromones(t+1) = pheromones(t) + r_1 * (pheromones_{min} - pheromones(t))$
Pour ramassage $pheromones(t+1) = pheromones(t) + r_2 * (pheromones_{max} - pheromones(t))$

Ce simple mécanisme de régulation assure qu'aucun agent ne sortira des limites physiologiques. Cependant, la vraie valeur de la quantité de pheromones déposée ou ramassée à chaque cycle de la simulation dépend de l'état actuel de l'agent : un agent avec une grande quantité de pheromone déposera plus et ramassera moins qu'un agent avec peu de pheromones. Cet arrangement permet aux chemins de pheromone à être vite construits (puisque les agents déposent plus de pheromones au départ) et disparaître rapidement (puisque les agents vers la fin de la tâche possèdent statistiquement peu de pheromones, alors ils ont tendance à en ramasser plus). Dans les simulations effectuées nous avons fixé $pheromones_{min}$=10 et $pheromones_{max}$ = 100, pour tous les agents.

3.3. Adaptation à deux niveaux

Un taux élévé de déposition sera de bénéfice au début et au milieu de la tâche, quand tous les agents veulent créer et renforcer un chemin rapidement, tandis qu'un taux élévé de ramassage sera de bénéfice à la fin de la tâche, quand les agents veulent détruire aussitôt que possible le

chemin à la source épuisée. Puisque ces paramètres sont contradictoires, nous aimerions introduire un mécanisme d'adaptation qui "reconnaîtrait" si l'agent se trouve vers le début ou vers la fin de la tâche et mettrait à jour les taux d'adaptation de premier niveau comme il convient. Pour ce faire, il faut un critère de "reconnaissance" d'état d'avancement de la tâche. Le seul critère sous la main puisse être la quantité de pheromone dans l'environnement. Puisque cette quantité globale n'est pas mesurable par les agents, nous utilisons une mesure d'estimation, à voir la quantité de pheromone sur la position de l'agent. Cette mesure est utilisée comme suit :

Pour déposition Si pheromones(t) ≥ estimation_environnement

$r_1(t+1) = r_1(t) + r_{r1} * (r_{1max} - r_1(t))$

sinon $r_1(t+1) = r_1(t) + r_{r1} * (r_{1min} - r_1(t))$

Pour ramassage *Si pheromones(t) ≥ estimation_environnement*

$r_2(t+1) = r_2(t) + r_{r2} * (r_{2min} - r_2(t))$

sinon $r_2(t+1) = r_2(t) + r_{r2} * (r_{2max} - r_2(t))$

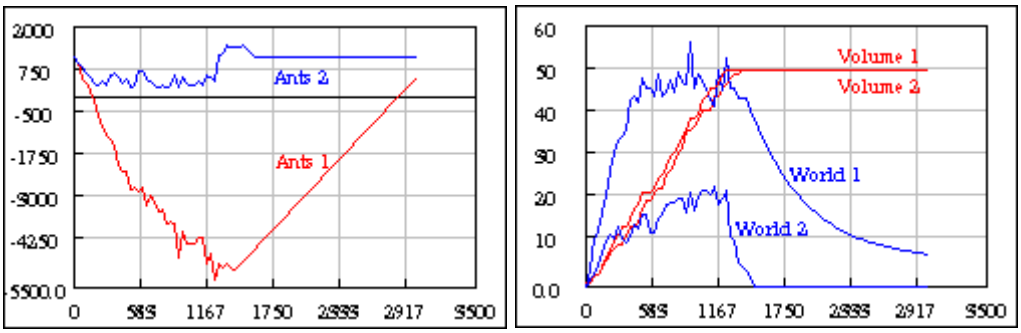

Figure 3. Comparaison de modèle de base (1) et modèle adaptatif (2) (gauche) Quantité de pheromones des agents, (droite) Quantité de pheromones dans l'environnement et volume cumulatif de transport au nid.

Ainsi, le taux de déposition de pheromones augmente quand le robot possède plus de pheromone qu'il perçoit dans son environnement et baisse autrement. Inversement, le taux de ramassage de pheromones augmente quand le robot possède moins de pheromone qu'il perçoit dans son environnement et augmente autrement. La figure 3 donne un résultat typique d'application du modèle précédent. A notre surprise, la régulation des taux de déposition et de ramassage modifie le comportement qualitatif des agents, puisque la quantité de pheromone dans le monde prend rapidement une valeur très élévée, reste près de cette valeur durant la tâche, et retombe à 0 très rapidement quand la source est épuisée, en démontrant pendant toute cette periode beaucoup moins de fluctuations que le modèle comportemental de base). Mais les résultats sont également affectés de façon quantitative, puisque la durée de la tâche (source épuisée et pheromones disparues) est bien inférieure à celle dans le modèle de base et celle dans le modèle avec adaptation à un seul niveau.

4 Étude de cas III : L'agent coopératif

4.1. Le problème

Une question majeure des recherches en vie artificielle et en biologie théorique est le comportement coopératif entre agents égoïstes. Le problème de coopération part de l'hypothèse que chaque agent a un intérêt personnel immédiat à défecter, tandis que le comportement

optimal joint serait de coopérer. Ce problème est traditionnellement modélisé comme un jeu de deux joueurs, le Dilemme Itéré des Prisonniers (Iterated Prisoner's Dilemma, IPD).

Ainsi, à chaque cycle d'interaction, les agents jouent au Dilemme du Prisonnier. Chaque agent peut coopérer (C) ou défecter (D). Les résultats obtenus pour chaque paire d'actions (agent,adversaire) sont donnés dans le tableau qui suit :

Agent	Adversaire	Résultat (score)
C	C	3 (= Récompense)
C	D	0 (= Stupide)
D	C	5 (= Tentation)
D	D	1 (= Punition)

Les expériences habituelles avec des stratégies pour l'IPD sont soit des tournois (championnats) soit des expériences écologiques. Dans les championnats, chaque stratégie joue contre toutes les autres et les résultats sont régroupés par stratégie à la fin. Dans les expériences écologiques, une population de stratégies joue en tournois et chaque génération successive regroupe les meilleures finalistes de la génération précédente à des proportions relatives à leurs scores.

La première stratégie importante pour l'IPD a été conçue et étudiée par Axelrod (1984); il s'agit de la stratégie "Tit For Tat" (abrégé TFT).

Commencer par coopérer,
Ensuite à chaque cycle d'interaction retourner la dernière action de l'adversaire.

Cette stratégie a obtenu les meilleurs résultats dans les premiers championnats et a été trouvée relativement stable dans des contextes écologiques.

Une des meilleures stratégies rencontrées dans la littérature est GRADUAL (Beaufils et al. 1996) qui obtient les meilleurs résultats face à quasiment toutes sortes de stratégies conçues. Cette stratégie commence en coopérant et ensuite joue TFT, sauf que sa défection n'est pas un seul D mais l'aveugle suite (nxD)CC, où n est le nombre des défections de l'adversaire dans le passé (mesure cumulative). Ainsi, GRADUAL répond avec DCC à la première défection de l'adversaire, DDCC à la deuxième, etc. La justification de ces performances est qu'elle punit de plus en plus l'adversaire, comme nécessaire, et ensuite l'apaise avec deux coopérations de suite.

Comme dans le cas de l'agent-fourmis, la première motivation pour notre modèle est la conviction que nous pouvons imaginer une stratégie comparable à GRADUAL quant à ses performances mais avec une mémoire qui ne soit pas irreversible et permanente. Au contraire, nous cherchons un modèle à base de TFT, mais plus adaptatif qui démontrerait gradualité comportementale et qui posséderait le potentiel de stabilité face à des mondes changeants (par exemple, lors de remplacements d'adversaire etc.).

Avant de procéder à la conception de notre modèle, nous avons étudié les résultats obtenus par GRADUAL. Les stratégies d'IPD rencontrées dans la littérature appartiennent à une des trois catégories suivantes :

- Des stratégies relativement complexes qui utilisent des informations de jeu; en général, il s'agit des stratégies coopératives jusqu'à ce que l'adversaire défecte, auquel cas elles adoptent une attitude de represailles (TFT, GRIM, GRADUAL, etc.).
- Des stratégies essentiellement coopératives mais qui commencent comme suspicieuses, par exemple un jouant quelque coups de D au départ, afin de tester leur adversaire. Cette famille de stratégies comprend entre autres la stratégie "suspicious tft" (STFT) et la stratégie "prober" de (Beaufils et al. 1996).
- Des stratégies qui sont clairement irrationnelles, parce qu'elles n'utilisent aucune information de jeu, mais elles suivent un rituel plus ou moins déterminé d'actions. Cette famille de stratégies comprend entre autres la stratégie aléatoire et toute stratégie aveugle périodique, telle que CCD, DDC etc.

Une stratégie maximisera ses résultats, si elle est capable de converger à un régime de coopération mutuelle face aux stratégies des deux premières catégories et à un régime de défection constante face aux stratégies de la troisième catégorie. Une défection constante face aux stratégies périodiques est nécessaire pour atteindre un score maximal (voir Tzafestas (2000) pour une preuve). La stratégie GRADUAL remplit toutes les deux conditions, car elle répond avec deux C consécutifs après une longue série de défections –ce qui donne la possibilité à STFT et les stratégies "prober" de reprendre coopération- et converge à ALLD face aux stratégies irrationnelles. Une solution du problème de la mémoire permanente doit démontrer les mêmes propriétés.

4.2. La solution : Reformulation du problème comme un problème d'adaptation

La stratégie adaptative que nous cherchons doit être essentiellement tit-for-tat. En plus, elle doit montrer moins d'oscillations comportementales entre C et D. Pour ce faire, la stratégie doit avoir une estimation de la stratégie de l'adversaire, coopérante ou irrationnelle, et y réagir à la tit-for-tat. L'estimation doit être mise-à-jour continûment durant l'interaction avec l'adversaire. Cela peut être modélisé à l'aide d'une simple variable continue, "l'image du monde", qui prend une valeur de 0 (défection parfaite) à 1 (coopération parfaite). Les valeurs intermédiaires représentent des degrés de coopération ou de défection. Le modèle tit-for-tat adaptatif peut alors être formulé comme un simple modèle linéaire :

Tit-for-tat adaptatif
Si (l'adversaire joua un C pendant le dernier cycle) alors
monde = monde + r(1-monde), r est le taux d'adaptation*
sinon monde = monde + r(0-monde)*
Si (monde ≥ 0.5) jouer C, sinon jouer D

Le modèle tit-for-tat habituel correspond au cas r=1 (adaptation immédiate à la dernière action de l'adversaire). Bien evidemment, l'utilisation de valeurs r<1 permettra l'émergence d'un comportement plus graduel et plus robuste aux perturbations. Les performances du modèle tit-for-tat adaptatif face aux trois types de stratégies présentés ci-dessus sont les suivantes :

- Face aux stratégies complexes qui utilisent des informations de jeu, le modèle coopère constamment et converge vite à une coopération parfaite.

- Face aux stratégies “suspicieuses”, le modèle joue exactement comme TFT, et la valeur de son “image de monde” oscille autour de la valeur critique de 0.5. (voir fig. 4 face à STFT).
- Face aux stratégies irrationnelles et périodiques, la valeur de son “image de monde” converge rapidement à des oscillations autour de la valeur critique de “nombre_moyen_de_C/ nombre_moyen_de_D” de l’adversaire.

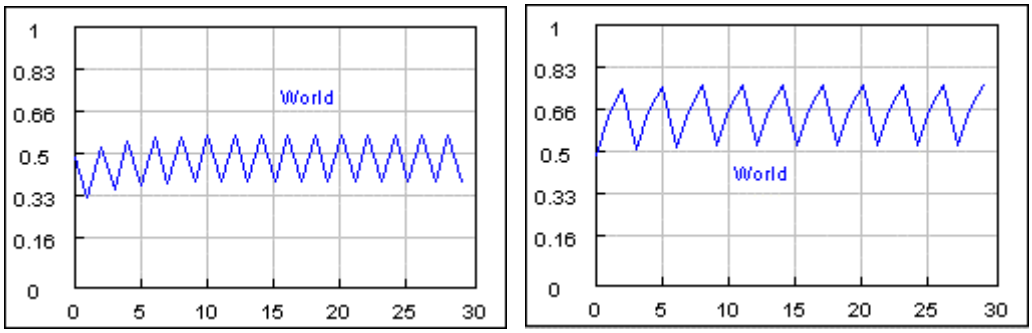

Figure 4. Comportement de tit-for-tat adaptatif (à gauche) face à STFT (r=0.2, monde(0)=0.5) (à droite) face à CCD ((r=0.2, monde(0)=0.5).

4.3. Adaptation à deux niveaux

La version précédente du modèle souffre de manipulabilité de sa variable “monde” par l’adversaire. Ceci se manifeste comme stabilisation de l’agent à un comportement oscillatoire (comme face à STFT) ou à un comportement parfaitement cooperatif face à des agents irrationnels (comme face à CCD). Pour palier à ce défaut, nous avons exploité notre observation que des taux différents de coopération et de défection (respectivement, r_c et r_d) conduiraient à des résultats différents. Plus spécifiquement, nous avons observé que l’agent tit-for-tat adaptatif parvient à amener les agents tels que STFT ou le “prober” à coopérer avec lui si $r_c>r_d$, tandis qu’il parvient à exploiter au maximum les agents irrationnels en devenant défectif si $r_c<r_d$. Ainsi, nous avons besoin d’un critère qui permettra à l’agent tit-for-tat adaptatif de découvrir si l’adversaire utilise une stratégie de représailles ou s’il est tout simplement irrationnel, afin d’adopter le papamétrage convenable. La loi d’adaptation qui a été développée est la suivante :

Pendant une fenêtre d’observation (window=w), enregistrer combien des fois (n) l’action de l’agent a coïncidé avec celle de l’adversaire. Dans des intervalles reguliers (tous les w cycles) adapter les taux comme suit :

Si (n>seuil) alors $r_c = r_{min}$, $r_d = r_{max}$, *sinon* $r_c = r_{max}$, $r_d = r_{min}$

Cette loi peut être interprétée comme :

*Si (le mode est assez coopératif)** *alors* $r_c = r_{min}$, $r_d = r_{max}$, *sinon* $r_c = r_{max}$, $r_d = r_{min}$

(*) *à rappeler que “mon action = l’action de l’adversaire” est le critère dit “pavlovien” de coopération (Nowak et Sigmund 1993)*

A noter enfin que l’agent converge à un taux de coopération faible quand son monde est trouvé cooperatif, sinon à un taux de coopération important, c’est-à-dire, il utilise des rétroactions négatives au niveau de régulation de taux. Ce modèle avec de l’adaptation à deux niveaux, parvient à obtenir la meilleure performance contre les agents de tous les types, tout en possédant une mémoire en principe reversible.

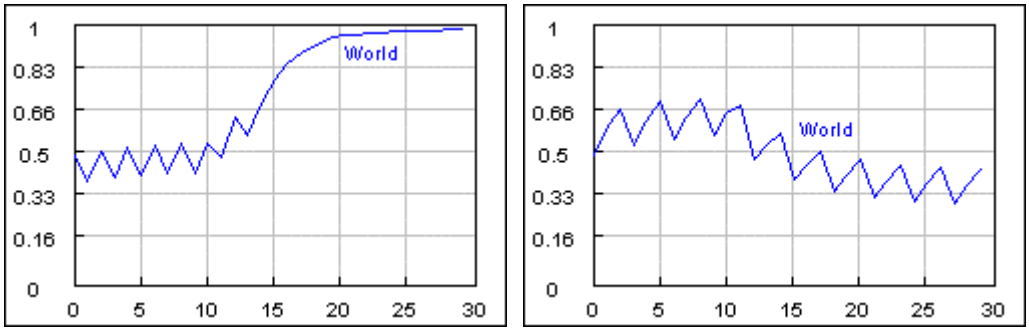

Figure 5. Comportement de tit-for-tat adaptatif méta-régulé (à gauche) face à STFT (à droite) face à CCD. Dans les deux cas $r_c(0)=0.2$, $r_d(0)=0.2$, $r_{max}=0.3$, $r_{min}=0.1$, monde(0)=0.5, w=10, seuil=2.

5 Adaptation, comportement, évolution

Dans tous les trois cas présentés auparavant, nous avons pu montrer que le comportement de l'agent dépend d'une variable critique qui dirige sa motivation pour agir. Cette variable est en réalité couplée avec l'environnement par l'intermédiaire même du comportement de l'agent. En essayant de réguler sa propre variable, l'agent essaye en effet de manipuler la variable correspondante globale. De plus, cette variable a une *valeur cognitive*, puisqu'elle représente l'idée que l'agent a pour l'état de son environnement. Sous cet angle de vue, l'agent essaye d'approximer son environnement, c'est-à-dire de s'adapter à lui. Les variables régulées sont dans ce sens critiques pour la "survie" ou l'opérationalité de l'agent et correspondent ainsi aux *variables essentielles* de Ashby (1960).

L'operationalité du comportement est assurée par le mécanisme d'auto-régulation de deuxième niveau, qui agit sur les taux d'adaptation du premier niveau. Cette approche est donc compatible avec l'approche dynamique de la cognition (van Gelder et Port 1995) qui déclare la dynamique du comportement comme un facteur primordial des capacités cognitives.

En somme, nous avons montré que l'opérationalité de l'agent dans un ensemble de problèmes est assurée grâce à un mécanisme de double régulation qui définit un système d'adaptation homéostatique de base, et un système de méta-adaptation qui agit sur les paramètres d'adaptation du niveau précédent. Notre perspective à long terme est de formaliser une théorie de régulation pour les agents autonomes. La notion d'adaptation qui agit sur elle même, c'est-à-dire d'adaptation réflexive, semble jusqu'à maintenant un moyen puissant de modélisation qui est en ligne avec les recherches en régulation classique Pour avancer sur cette voie, il faudra ultérieurement répondre à un nombre de questions :

- Comment peut-on identifier la variable cognitive critique pour chaque problème ? Et comment formalise-t-on la régulation (de premier niveau) pour chaque problème ? En effet, quels sont les problèmes qui peuvent être étudiés de cette façon ?
- Combien de taux de premier niveau faut-il ? Ou, combien de processus indépendants de régulation existe-t-il ? Par exemple, l'agent explorateur possède un taux d'adaptation, tandis que les deux autres agents en possèdent deux.
- Quel est le critère de méta-régulation ? Dans tous les trois cas étudiés, le critère est purement qualitatif et dépendant du problème. Peut-on trouver une loi générale conceptuelle qui regroupe tous les cas ?

- Quelle est la nature des dynamiques de méta-régulation ? Dans tous les trois cas, nous avons montré (Tzafestas 2000, soumis 1, 2) qu'une dynamique "bang-bang" (haut-bas) est suffisante, parce que ce qui compte est la relation entre taux et dynamiques et non pas les valeurs absolues des paramètres.
- Finalement, quel est le rôle et la valeur du "comportement dans le vide" (sans perturbation) ? Ce comportement est purement spécifique à l'agent en question et peut être différent d'agent à agent, à cause des paramétrages différés qui définissent ainsi la "personnalité" de chaque agent. Des expériences dans le vide (Tzafestas soumis 1, 2) montrent que l'étude dans le vide permet certaines prédictions de performance de l'agent.

Mais le mécanisme d'adaptation réflexive n'est pas simplement le contrôleur du comportement; il peut servir également de structure évolutive. Selon cette perspective, le "code génétique" d'un agent que nous faisons évoluer n'est pas un ensemble de paramètres numériques auquels l'évolution donnera des valeurs qui "optimisent" les performances de l'agent, mais l'ensemble de critères du mécanisme d'adaptation réflexive auquels l'évolution donnera une signification en les "branchant" directement aux capteurs de l'agent. Dans notre cas, le code génétique consiste à l'ensemble structurel suivant [perception de la variable du niveau 1, critère de divergence du niveau 2]. L'instantiation de cette structure pour les trois agents étudiés est comme suit :

	Perception de variable du niveau 1	***Critère de divergence du niveau 2***
Agent explorateur	diff = p_{calc} – var p_{calc} = nombre des échantillons ramassés / nombre des pas effectués	|diff| (= |p_{calc}-var|) ≤ f_p
Agent-fourmis	*Si déposition, diff = var_{min} – var* *Si ramassage, diff = var_{max} – var*	(var - estimation_environnement) ≥ 0
Agent TFT adaptatif	*Si (l'adversaire joua un C pendant le dernier cycle)* diff = 1-monde *Sinon* diff = 0-monde = -monde	n≤ seuil, où n = combien des fois l'action de l'agent a coïncidé avec celle de l'adversaire

En effet, la formulation de ce tableau est en liaison directe avec les questions posées auparavant, car les deux premières questions concernent la perception de la variable du niveau 1, tandis que la troisième question concerne le critère de divergence du niveau 2. L'analogie conceptuelle des trois mécanismes nous oblige à en disséquer les différences (nombre de taux, directions de régulation, dynamique bang-bang ou graduelle). Dans le cas idéal nous aimerions avoir un seul taux d'adaptation et une seule loi d'effet pour chaque niveau, comparaison à 0 et dynamique bang-bang dans le deuxième niveau. Il faut souligner ici que le mécanisme restera invariant à travers les générations, mais les structures utilisées changeront. Ceci permettra la "reprogrammation" évolutive des agents pour répondre aux besoins dynamiques d'adaptation à un environnement évolutif et notamment à un environnement dans lequel les agents co-évoluent continûment.

Le deuxième niveau d'adaptation formulé ainsi permettrait le passage rapide à une modélisation génétique, car on sait que les vrais gènes prennent des valeurs booléennes et régulent des vitesses de réactions chimiques à l'intérieur des organismes biologiques. Nous imaginons que cette perspective d'adaptation à deux niveaux permettrait la modélisation simplifiée d'un nombre de comportements biologiques complexes de tout niveau. Nous pensons plus particulièrement à des comportements tels que l'auto-organisation des réseaux de gènes (Raeymaekers 2002), le développement et la stabilisation des communications cellulaires

(Furusawa et Kaneko 1998), la répartition de tâches dans une société animale (Ratnieks et Anderson 1999), certains effets de spéciation (Dieckmann et Doebeli 1999) etc.

Bibliographie

Ashby, W.R. (1960). *Design for a brain - The origin of adaptive behaviour*. 2nd edition, Chapman & Hall.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Beaufils, B., Delahaye, J.-P. et Mathieu, P. (1996). "Our meeting with gradual: A good strategy for the iterated prisoner's dilemma", *Proceedings Artificial Life V*, Nara, Japan.

Beckers, R., Holland O.E. et Deneubourg, J.-L. (1994). "From local actions to global tasks: Stigmergy and collective robotics", *Proceedings Artificial Life IV* (R. Brooks and P. Maes, Eds.), MIT Press, 181-189.

Deneubourg, J.-L., Aron, S., Goss, S. et Pasteels, J.M. (1990). "The self-organizing exploratory pattern of the Argentine Ant", *Journal of Insect Behavior* 3:159-168.

Dieckmann, U., Doebeli, M. (1999). On the origin of species by sympatric speciation, *Nature*, 400(6742):311-2, 22 July.

Drogoul, A. et Ferber, J. (1992). "From Tom Thumb to the Dockers : Some experiments with foraging robots", *Proceedings SAB 1992*, 451-459.

Furusawa, C. et Kaneko, K. (1998). Emergence of rules in cell society : Differentiation, Hierarchy and Stability, *Bulletin of Mathematical Biology*, 60:659-687.

Korzeniewski, B. (2001). Cybernetic formulation of the definition of life, *Journal of Theoretical Biology*, 209:275-286.

Mataric, M. (1992). "Designing emergent behaviors : From local interactions to collective intelligence", *Proceedings SAB 1992*, 432-441.

Nowak, M. et Sigmund, K. (1993). "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game", *Nature* 364(1993):56-58.

Raeymaekers, L. (2002). Dynamics of boolean networks controlled by biologically meaningful functions, *Journal of Theoretical Biology*, 218:331-341.

Ratnieks, F.L.W., Anderson, C. (1999). Task partitioning in insect societies, *Insectes Sociaux*, 46:95-108.

Solé, R.V., Bonabeau, E., Delgado, J., Fernández, P. et Marín, J. (2000). Pattern formation and optimization in army ant raids, *Artificial Life*, 6(3): 219-26.

Steels, L. (1990). "Towards a theory of emergent functionality", *Proceedings SAB 1990*, 451-461.

Tzafestas, E. (1995). *Vers une systémique des agents autonomes : Des cellules, des motivations et des perturbations*, Thèse de Doctorat de l'Université Pierre et Marie Curie, Paris, décembre 1995.

Tzafestas, E. (1998). "Tom Thumb Robots Revisited : Self-Regulation as the Basis of Behavior", *Proceedings Artificial Life VI*, San Diego, CA.

Tzafestas, E. (2000). "Toward Adaptive Cooperative Behavior", *Proceedings SAB 2000*, Paris, France.

Tzafestas, E. (soumis 1). "Regulation Problems in Explorer Agents", submitted.

Tzafestas, E. (soumis 2). "Pheromone regulation in insect foraging", submitted.

van Gelder, T. et Port, R. (1995). "It's about time : An overview of the dynamical approach to cognition", in *Mind as Motion : Explorations in the dynamics of cognition*, by T. van Gelder and R. Port, MIT Press.