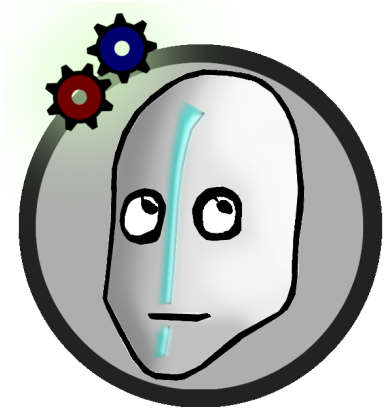

ACRONYME	ROBOERGOSUM
NOM DU PROJET	ROBOTS CONSCIENTS
REFERENCE	DECISION ANR-12-CORD-0030
NUMERO DE LA TACHE	T4
NOM DE LA TACHE	Learning abilities
NUMERO DU RAPPORT	D4.3
TITRE DU RAPPORT	Report on decision criteria for switching strategies, goals and meta-goals based on frustration measures
PARTENAIRES	ISIR, LAAS
DATE	Juin 2014

1. RESUME DU DELIVERABLE / SUMMARY

Dans l'approche proposée dans la tâche 4, nous souhaitons que le robot puisse détecter les changements de situations, afin d'adapter son comportement au mieux à l'aide des experts dont il dispose. Le comportement est une réponse directe aux buts que le robot suit. Pour obtenir un agent totalement autonome, ces buts doivent être fixés par le robot lui-même, en fonction des requêtes de l'environnement et de son état interne.
Au niveau de la sélection de l'action, on sépare les informations reçues par le système en informations externes et informations internes. Les informations externes proviennent de la perception, et servent à construire l'état manipulé par les algorithmes d'apprentissage par renforcement. Les informations internes viennent des experts eux-mêmes ou des autres systèmes du robot avec lesquels la sélection de l'action communique. Nous nous intéressons plus particulièrement aux informations internes, puisque nous souhaitons alterner entre les stratégies indépendamment de la nature de tâche.

2. APPROCHE

Les modèles d'apprentissage des habitudes dont nous nous inspirons (Keramati et al., 2011), (Dezfouli et Balleine, 2012), (Caluwaerts et al., 2012) proposent des critères et manières d'alterner entre les experts. Les deux premiers modèles s'appuient sur un critère bien défini, tandis que le modèle de (Caluwaerts et al., 2012) apprend au niveau du Meta-Contrôleur quel expert est le plus pertinent dans un état donné. Par ailleurs, Keramati et al. utilise le critère au niveau de la décision (choix d'agir en fonction des connaissances actuelles ou d'affiner ces connaissances) tandis que Dezfouli et Balleine s'en sert pour construire ou non l'habitude.

Dans ces deux cas, le critère fait intervenir la récompense moyenne reçue par le système[1]. Nous considérons la récompense moyenne comme un signal interne venant d'un système de motivation du robot, même si ce signal est souvent donné par une source extérieure (environnement, simulateur, « professeur » qui guide l'agent dans la tâche). En effet, cette notion de récompense est relative : la même source extérieure (e.g. la présence de nourriture pour un animal) peut avoir des valeurs différentes selon l'état courant des motivations de l'agent (e.g. un animal à satiété n'est plus intéressé par la nourriture). Nous nous intéressons donc à l'information qui peut être extraite de ce signal de récompense moyenne :

- une récompense moyenne constante indique que le comportement de l'agent a trouvé un comportement stable en terme de récompense dans l'environnement. L'intérêt du comportement en question est mesuré par la valeur de la récompense moyenne. La

[1] Calculée selon une fenêtre glissante

difficulté de la tâche, qui oppose les récompenses et coûts reçus, influe également sur l'amplitude des valeurs de la récompense moyenne, à environnement et tâche constants.

- une récompense moyenne croissante indique soit que le comportement de l'agent est en cours d'apprentissage et s'améliore, soit que l'environnement change de sorte à ce que le comportement suivi devienne meilleur. Nous pouvons donc extraire une information soit sur l'évolution du comportement (l'agent est en train d'apprendre) soit sur l'environnement (qui est en cours de modification).

- une récompense moyenne décroissante indique soit que le comportement de l'agent est en cours d'apprentissage et fait des mauvais choix, soit que l'environnement a changé et le comportement de l'agent n'est plus adapté pour résoudre la tâche.

Un premier critère très simple qui utilise ce signal de récompense est le suivant :
si la récompense moyenne croît ou est constante, on donne la main à l'Expert habituel. Si elle décroît, on donne la main à l'Expert dirigé vers un but.

Ce critère permet de céder la main à l'Expert habituel dès lors qu'un comportement performant est trouvé, et donc de ne pas perdre de temps à planifier, et de revenir à l'Expert dirigé vers un but lorsque les choix ne rapportent pas assez de récompense ou que l'environnement a changé. Ce critère s'appuie sur l'hypothèse que l'Expert dirigé vers un but a une information bien meilleure que l'Expert habituel et trouvera plus vite un nouveau comportement adapté aux nouvelles conditions de l'environnement.

En pratique, cette hypothèse n'est pas vraie sans un mécanisme visant à guider la recherche dans l'Expert dirigé vers un but. Lorsque cet Expert apprend le modèle de Transition et de Récompense entre états et actions d'une tâche réaliste, le nombre d'états générés produit une explosion combinatoire qui accroît le temps nécessaire à la planification. Par conséquent, l'Expert ne peut évaluer complétement les conséquences de ses actions, et conserve une information imprécise, ce qui invalide l'hypothèse « d'information parfaite ». La récompense moyenne reste cependant un signal important pour évaluer l'adaptation de notre comportement vis-à-vis de la tâche actuelle.

Une autre information importante est la distribution de probabilité de chaque Expert lorsqu'il rend sa décision. Cette information est utilisée notamment dans les méthodes dites « d'*Ensemble Reinforcement Learning* » (Wiering & Hasselt, 2008) et s'appuie sur le fait que nos Experts décident de manière probabiliste quelle action accomplir. La littérature propose quatre méthodes pour choisir ou fusionner les actions, nous proposons une cinquième mesure qui apporte également une certaine information sur l'état des Experts lors de leur décision.

Wiering & Hassel propose de générer une distribution de probabilité « finale » à l'aide d'une distribution de Boltzmann sur des valeurs de préférence. Ces valeurs de préférence sont calculées selon quatre méthodes :

- Vote majoritaire : l'action la plus probable pour chaque Expert reçoit une valeur de 1.0, les autres 0.0. La valeur de préférence finale pour une action est la somme de sa valeur pour chaque Expert. À noter que dans notre implémentation, n actions partageant la plus haute probabilité reçoivent 1/n pour représenter le fait que le choix est moins contrasté que si une seule action est la meilleure.

- Vote selon le rang : les n actions de chaque Expert sont ordonnées par leur probabilité d'être choisie, la première reçoit un score de n, la seconde n-1, etc. Deux actions de même probabilité reçoivent le même score. La préférence est la somme sur les Experts des scores de chaque action.

- Multiplication de Boltzmann : chaque action voit sa probabilité multipliée sur les Experts.
- Addition de Boltzmann : chaque action voit sa probabilité sommée sur les Experts, cette méthode est proche du Vote selon le rang en conservant la finesse des probabilités des actions.

Ces critères ont été implémentés sur le Meta-Contrôleur et sont en cours de test.

Par ailleurs, nous proposons une mesure sur les probabilités légèrement différente : nous calculons l'entropie de la distribution de probabilités d'action d'un Expert, ce qui nous donne un score qui décrit le contraste dans la distribution, et estime la « confiance » de chaque Expert dans son choix. Le critère est le suivant :

le contrôle est donné à l'Expert le plus confiant i.e. celui de plus faible entropie.

Ce critère permet de donner la main à l'Expert habituel dès lors qu'il a appris suffisamment pour être plus confiant que l'Expert dirigé vers un but (qui reste assez incertain étant donné l'hypothèse non vérifiée « d'information parfaite »). Le défaut de cette mesure est que l'Expert habituel met du temps à se réadapter à de nouvelles conditions environnementales, et conserve dont une entropie faible malgré l'inadéquation de la stratégie qu'il propose. L'entropie indique donc lequel des Experts est le plus certain de ses choix mais ne tient pas compte des évolutions de l'environnement.

Une dernière information utilisable est la variation des Q-valeurs de l'Expert habituel, indicatrice du degré d'apprentissage de l'Expert. Il n'existe pas de mesure strictement équivalente pour l'Expert dirigé vers un but, qui apprend les fonctions de Transition et de Récompense, mais on peut obtenir une estimation de la convergence de la planification en

calculant la variation des Q-valeurs planifiées. Ces mesures n'ont pas été testées au sein d'un critère mais leur intégration et leur étude pratique sont prévues.
La mesure de frustration proposée dans (Hasson & Gaussier, 2010) apporte un mécanisme intéressant d'auto- évaluation du comportement. Cependant, il fait un certain nombre d'hypothèses qui le rendent difficilement applicable au niveau des stratégies :

- Le but est défini : la notion de but à atteindre, dans les algorithmes d'apprentissage par renforcement, est floue. L'agent vise à maximiser la récompense reçue, ce qui se traduit par trouver le chemin qui permet d'accumuler le plus de récompense au cours de la vie de l'agent. Il peut exister un état final au sens où l'état atteint et l'action effectuée permettent à l'agent de rester dans cet état (e.g. un état qui correspond à la position finale et une action « rester » sur place) de même que la tâche peut être répétitive et donc sans fin (pousser des blocs d'un tapis roulant dans un bac et recevoir une récompense maximale lorsque le bloc tombe dans le bon bac amène dans un état où un bloc a été enlevé mais de nouveaux arrivent). Par ailleurs, cette notion de but n'est trouvée que dans la fonction de Récompense de l'Expert dirigé vers un but, l'Expert habituel cachant la récompense dans les associations état-action.

- Le but est connu : nos algorithmes ne connaissent pas forcément le but à atteindre avant d'avoir exploré l'environnement et découvert un couple état-action qui rapporte une récompense élevée. Tant que ce couple n'a pas été découvert, calculer une distance au but n'est pas possible, et on ne peut mesurer de frustration pour détecter un comportement qui ne progresse pas. Cette mesure est cependant pertinente à une niveau plus élevé, dans le système de motivation : en effet, c'est ce système qui détermine les buts à atteindre par l'agent en fonction des différents besoins. Il est donc possible de mesurer l'évolution de la distance de l'agent au but, et de renvoyer une information de frustration vers le système de décision de l'action, qui pourra donc faire évoluer sa stratégie en conséquence.

En conclusion, nous avons étudié un certain nombre de signaux et de critères simples. Afin d'avoir un comportement plus optimal, il ressort que l'arbitrage entre les Experts doit tenir compte de tous ces signaux afin de sélectionner la bonne stratégie au bon moment. Par ailleurs, l'interaction avec les niveaux supérieurs doit également permettre l'utilisation d'informations non normalement accessibles au système de décision afin que l'agent soit capable d'auto-évaluer son comportement.

3. REFERENCES

(Keramati et al., 2011) : Mehdi Keramati, Amir Dezfouli, Payam Piray (2011), *Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes*, In PLoS Computational Biology, Vol. 7, No. 5., doi:10.1371/journal.pcbi.1002055

(Dezfouli et Balleine, 2012) : Dezfouli, A. and Balleine, B. W. (2012), *Habits, action sequences and reinforcement learning. European Journal of Neuroscience*, 35: 1036–1051. doi: 10.1111/j.1460-9568.2012.08050.x

(Caluwaerts et al., 2012) : Ken Caluwaerts, Antoine Favre-Félix, Mariacarla Staffa, Christophe Grand, Benoît Girard, Mehdi Khamassi (2012), *Neuro-inspired Navigation Strategies Shifting for Robots: Integration of a Multiple Landmark Taxon Strategy*, Biomimetic and Biohybrid Systems - First International Conference, Living Machines 2012, Barcelona, Spain, July 9-12, Proceedings, 62-73, doi :10.1007/978-3-642-31525-1_6

(Wiering & Hasselt, 2008) : Marco Wiering and Hado van Hasselt (2008). *Ensemble Algorithms in Reinforcement Learning*. In: IEEE Transactions, SMC Part B, special issue on Adaptive Dynamic Programming and Reinforcement Learning in Feedback Control. August 2008. pp. 930-936.

(Hasson & Gaussier, 2010) : Cyril Hasson & Philippe Gaussier (2010), *Frustration as a generical regulatory mechanism for motivated navigation*, in IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan, 4704—4709