# Learning Epistemic Actions in Model-Free Memory-Free Reinforcement Learning: Experiments with a Neuro-robotic Model

Dimitri Ognibene[1], Nicola Catenacci Volpi[4],
Giovanni Pezzulo[2,3], and Gianluca Baldassare[2,⋆]

[1] Personal Robotics Laboratory, Imperial College London, UK
[2] Istituto di Scienze e Tecnologie della Cognizione, CNR, Italy
[3] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Italy
[4] IMT Institute for Advanced Studies, Lucca, Italy

**Abstract.** Passive sensory processing is often insufficient to guide biological organisms in complex environments. Rather, behaviourally relevant information can be accessed by performing so-called *epistemic actions* that explicitly aim at unveiling hidden information. However, it is still unclear how an autonomous agent can learn epistemic actions and how it can use them adaptively. In this work, we propose a definition of epistemic actions for POMDPs that derive from their characterizations in cognitive science and classical planning literature. We give theoretical insights about how partial observability and epistemic actions can affect the learning process and performance in the extreme conditions of model-free and memory-free reinforcement learning where hidden information cannot be represented. We finally investigate these concepts using an integrated eye-arm neural architecture for robot control, which can use its effectors to execute epistemic actions and can exploit the actively gathered information to efficiently accomplish a seek-and-reach task.

## 1 Introduction

When an agent is executing a task in a non-deterministic and partially observable environment its behavior is affected by its limited knowledge. Recent evidence in neuroscience [1–3] indicates that living organisms can take into consideration the confidence in their knowledge and execute actions that allow the decrease of uncertainty if they satisfy a value/cost trade-off. These actions are named *epistemic actions* in cognitive science and in the planning literature, and *information-gathering actions* in operation research.

In robotics, epistemic actions have been applied in several tasks such as navigation (e.g., moving to positions where sensors can perceive to landmarks [4, 5]), active vision (e.g. moving the camera to acquire information given the limited

[⋆] This research was funded by the European Projects EFAA (G.A. FP7-ICT-270490), Goal-Leaders (G.A. FP7-ICT-270108), and IM-CLeVeR (G.A. FP7-ICT-IP-231722).

N.F. Lepora et al. (Eds.): Living Machines 2013, LNAI 8064, pp. 191–203, 2013.

field of view, the occlusions, and the changes in the environment [6]), tactile exploration [7, 8], and the active use of bio-inspired sensors such as rat whiskers [9]. The intrinsic complexity of the real world, however, may require even more versatile strategies like the execution of actions that change the environment in order to facilitate perception. Some examples are: opening a box or a drawer to see its content; rotating a picture to inspect its back; digging the ground to find root crops or moving the foliage to find fruits. Most of these actions cannot be predefined in the same way sensor controls are, because they are not purely epistemic. Indeed, in addition to changing the agent knowledge they also change the state of the environment and as a consequence they might result to be maladaptive for the agent.

A typical approach to solve the lack of information is using memory of previous perceptions. However, in some situations acting without appropriate knowledge is not useful (e.g., trying to open a safe with a limited number of attempts) so the first actions to execute should be directed to gather information (e.g., asking the opening number). Acting ignoring ignorance is seldom a good strategy. However, it is also not easy to devise at design time which actions an agent should execute to decrease uncertainty or which hidden structure of the environment it will encounter.

## 2 Epistemic Actions and POMDP Approximations

In the classical AI planning literature the problem of limited knowledge has been faced by adding knowledge preconditions to classical action definitions and through the definition of *epistemic actions* [10]. Knowledge preconditions define "what" information must be acquired, and the epistemic actions define "how". A common characteristic is that epistemic actions change only the agent knowledge of the world, and, differently from *pragmatic* or *ontic* actions, they do not change the world state [10, 11]. However, as we discussed earlier some actions affect both the perception (and knowledge) of the agent and the state of the environment. Given this ambiguity, a common choice is to model an action as a combination of an ontic action and an epistemic action [11].

In cognitive science epistemic actions are actions executed by a bounded agent as ways to overcome its intrinsic limits [12, 13]. When the limits being tackled are of perceptual nature we can use the expression *external epistemic actions* because they can be easily defined, once known the perceptual apparatus of the agent and the structure of the environment, without using information of the internal structure of the agent.

POMDPs [5] formalise the problem of optimising sequential decisions in partially observable stochastic environments. Agents have a complete probabilistic model of the environment, composed by a set of states $\mathcal{S}$ (and the related transition probabilities $\tau(s^{t+1}, a_t, s^t)$), and a set of observations (and again the related transition probabilities). At each step an agent executes an action $a$ and receives an observation $o$ and a reward $r$, finally it updates a probabilistic distribution of the state $s$, named *belief state* $b(s)$, (for which transition probabilities

$\tau(b_{t+1}, a_t, b_t)$ can be computed). An optimal behaviour is associated with an optimal value function $V(b)^* = E[\sum_{t=1}^{\infty} \gamma^t R_t | B = b]$. When the full state is observable the optimal behaviour and related value function can simply use the state $V(s)^* = E[\sum_{t=1}^{\infty} \gamma^t R_t | S_0 = s]$.

Classical POMDP theory does not make any explicit difference between epistemic and ontic actions. We propose a definition of *external epistemic actions* for the POMDP framework that takes inspiration from its definition in cognitive science and the classical planning literature. A starting point for the definition are the two main characteristics that Herzig and colleagues [11] associate to epistemic actions: *informativeness* and *non-intrusiveness.*

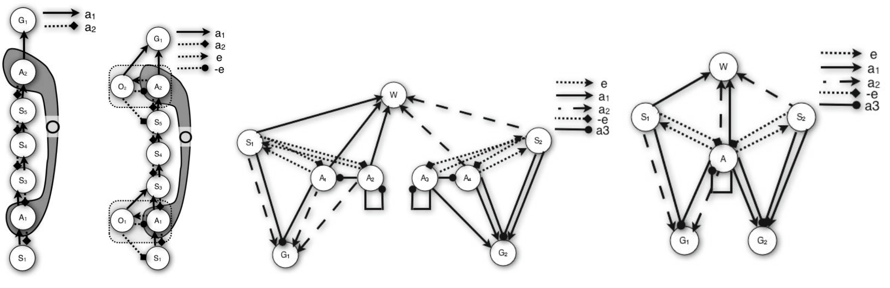
An action is *informative* when its execution reduces the uncertainty of the belief state $b(s)$. Formal definitions can be found in [5, 14]. More recent approaches, using non myopic value of information, can be found in [4, 15]. We define an action $e$ to be *non-intrusive* in the belief state $b$ if the expected value of the belief state reached after executing $e$, $E[Q(b, e)] = \sum_{b'} \tau(b, e, b') \sum_s b'(s) V^*(s)$, is equal to the value of the current belief state $b$ computed using the Q-MDP approximation $V^*(b) = \sum_s b(s) V^*(s)$ and if the immediate expected reward of executing the action is 0. $(\sum_s |b(s) r(s, e)| = 0)$[1]. The use of $V^*(s)$ in place of $V^*(b)$ is intuitively explainable by the fact that the latter comprises the modification to the internal state of the agent, thus any epistemic action will affect it. This formalises the concept that the execution of a non-intrusive action does not change the reward that the agent can receive.

We define an action $e$ to be *strictly external epistemic* in the belief state $b$ if it is informative and non-intrusive in the belief state $b$. An action $e$ is *strictly epistemic over observation* $o$ if the action is epistemic for every belief state $b$ for which $P(o|b) \neq 0$. A POMDP is an *MDP-reducible-POMDP* when it admits a policy $\pi_e$ that for any belief state reduces state entropy to zero in a finite number of steps using only epistemic actions. A MDP-reducible-POMDP can be solved combining the policy $\pi_e$ with the optimal policy $\pi_{MDP}$. The obtained solution can be sub-optimal because of the time spent executing $\pi_e$. This condition can be found in real-world tasks. For example, in some active vision problems [16] the algorithms used are based on a phase in which only the point of view is changed till the agent has enough confidence on the observed state [17]. This is a subclass of the MDP-reducible-POMDPs because only camera control actions are executed for information retrieval and the state of the task is unchanged.

## 3 Epistemic Actions and Reinforcement Learning

In the hypothesis that an agent is working in POMDPs and is endowed with epistemic actions, or even in the more strict condition of a MDP-reducible-POMDP, the problem remains of acquiring a complete stochastic model of the environment before solving it, which can still be complex. A different approach is using a reinforcement learning (RL) based agent which directly learns the

[1] Note that this is different from the optimal transition which states that $V^*(s) = max_a(r(s, a) + \gamma \sum_{s'} \tau(s', a, s) V^*(s'))$.



(a) (b) (c) (d)

**Fig. 1.** (a,b) Examples of POMDPs with ambiguity and optimal memory-free policy not learnable through RL. Circles that fall in the same grey shape will give the same ambiguous observation, while circles inside the same rectangle are observations of the same state. The other circles are unambiguous states. (c,d) A POMDP and its underlying MDP problem. Its peculiarity is the presence of $a_3$ that always brings the agent to the goal from the aliased states $A_{1-4}$.

action policy by interacting with the environment. Studying how RL, especially model-free and memory-free RL [18, 19], performs when epistemic actions are available in POMDP is interesting for two main reasons:

1. Reinforcement learning has been shown to be able to learn in MDP, however it has also shown to have strong limitations when facing POMDPs [17].
2. Learning a model can allow the execution of epistemic actions related to the hidden states that are correctly represented, but it is a complex problem by itself and cannot help for not represented states.
3. Many greedy POMDP algorithms have been shown to fail in POMDP because they "ignore ignorance" even if they use a complete belief state, so how a model-free [2] reinforcement learning agent can take into consideration its own ignorance and select epistemic actions is an open issue.

Moreover, while it has been shown that in some conditions memory-free agents can effectively behave in partially observable environments [20], limited work has been done to allow the *autonomous development* of such behaviours with similar constraints (but see [18, 21]).

### 3.1 Learning in POMDP with and without Epistemic Actions

A typical POMDP, originally reported in [22], is shown in Figure 1.a. The states $A_1$ and $A_2$ are ambiguous because they result in the same observation $O$.

[2] Note that a model-free, memory-free agent is also belief-free and unaware of its uncertainty. To define epistemic actions for such an agent, an external probabilistic observer has to be used, which receives as input the agent actions and observations. The changes in the uncertainty in the belief state of the observer after each action-observation pair is used to measure the action informativeness.

The optimal policy would be to select the action $a_1$ in every state. The use of a value iteration algorithm with the assumption that observations are states will result in the agent oscillating around state $A_1$[3]. Online learning using the exploration approach proposed in [23] obtains the same results without ever learning a path toward the goal. See [24] for more examples of the effect of perceptual aliasing on RL. A similar situation can arise in the POMDP shown in Figure 1.c,d.

Does the addition of epistemic actions allow a model-free memory-free RL agent to learn in these environments? Looking the environment of Figure 1.b, where epistemic actions and the related observations $O_1$ and $O_2$ are added, we can see that the agent will still move from state $S_3$ toward state $A_1$. So in this case adding epistemic actions does not solve the problem.

Epistemic actions affect learning by increasing the distance between the starting state and the goal, and by increasing the fan out of the resulting graph in comparison to the underlying MDP (each observation is connected to all the states that make it visible). The value of executing an epistemic action depends on the value of all the possible states reached, in other words on the estimated value of the information acquired. This results in an increase of the exploration time during learning. To see this, consider an environment like the one in Figure 1.c,d, and an agent that is endowed with epistemic action $e$ and $n_A$ ontic actions available in all states which, if executed in the right MDP state, result in the reward $R$ (in the figure $a_1$,$a_2$ and $a_3$). Similarly to the previous situation, the agent will not be able to distinguish between observations and states, so the observation $A$ in the POMDP is seen as a single state and is shared by the MDP states $A_1$,$A_2$,$A_3$,$A_4$. Note that action $a_3$ if executed several times can take the agent to the solution from every state. During the first trials of learning the epistemic action $e$ will have very low probability of being executed. Consequently the (unambiguous) states, e.g. $S_1$, will not be evaluated correctly. In the unambiguous state the agent will have also very little probability of learning the right action to do because of the a high number ($n_A$) of ontic actions. Thus, after the first trials the epistemic actions will not probably increase their values. At the same time several ontic actions in the ambiguous state $A$ will be executed, and so there will be a high chance that one of them leads to a reward. This will result in increasing the probability of reselecting the same action while the probability of selecting an epistemic actions will decrease. Moreover, if there is an action like $a_3$ available in the ambiguous state which cycles through hidden states and also brings the agent to the goal, it will be easily found and its value will not decrease due to punishments. Even if the policy resulting from $a_3$ is sub-optimal it requires the exploration of small set of states and will slow down the exploration of other actions.

Given this kind of dynamics, before an epistemic action can acquire a high value the agent should learn how to behave in most of the non-ambiguos states

[3] We consider $O$ as a state with transition probabilities resulting from a different mixture of those of state $A_1$ and of state $A_2$. The value function obtained in observation $O$ has higher value than in state $S_4$, so in state $S_3$ the agent will choose action $a_2$ and not action $a_1$.

$S_i$. It is interesting to note that using a greedy method for POMDPs which ignores uncertainty reduction [5], the action chosen will be one of the ontic actions. With enough learning experience a RL agent might be able to learn a policy comprising epistemic actions because the learnt value of action-state pairs can be comprehensive of the information value.

## 4 Experimental Results

We used a neuro-robotic system of arm-eye coordination to study experimentally the concepts presented in the previous sections. The architecture of the model (Figure 2.a) integrates two components: (a) an attention control component formed by a bottom-up and a top-down attention sub-component; (b) an arm control component. Only an overview of these components is presented here. For a complete description of the system refer to [25][4].

The setup used to test the model is a simulated version of a real system presented in [24] (see Figure 2.a), formed by a down-looking camera and a 2-DOFs robotic arm. The arm horizontal working plane is formed by a horizontal computer screen where the task stimuli appear. The camera image activates a *periphery map* that implements bottom-up attention. The central part of the image (*fovea*) feeds a *reinforcement-learning actor-critic* [19] component (implemented by two feed-forward neural networks) that learns to predict the positions of relevant visual elements based on the currently foveated cues (top-down attention). A *saliency map* sums up the information from the periphery map with the output of the actor network and selects the next eye movement corresponding to the most active neurons (through neural competition). Each eye fixation point, encoded in a *eye posture map*, suggests a potential arm target to an *arm posture map* which (a) performs the "eye posture → arm posture" inverse kinematic and (b) implements a second neural competition which triggers reaching movements when the eye fixates the same location for about three consecutive time steps. If the reached target is the correct one (red object), the actor-critic component is rewarded. By closely coupling reaching to gaze control, the proposed model embodies the "attention-for-action" principle [26]. This principle states that in organisms attention has the function of extracting the information necessary to control action. This principle might be incorporated in the system as a hard-wired link between an epistemic (visual) action and an ontic (reaching) action. The following experiments will test if the model is able to learn to execute epistemic actions which do not have such a pre-wired link in the architecture.

[4] For this work three minor change where made to the architecture: (a) the foveal input was pre-processed in order to separate the different objects on different input units both for the actor and the critic neural networks: this informs the agent on the identity of objects; (b) the action of reaching is not directly punished to satisfy the non-intrusiveness constraint, but for every saccade the agent gets a punishment of -0.0025; (c) Finally PAM, an action-based memory system, was removed to have a model/memory/belief-free agent, so the agent could not rely on the inhibition of return mechanism and on previous estimation of target position, some key properties of PAM.

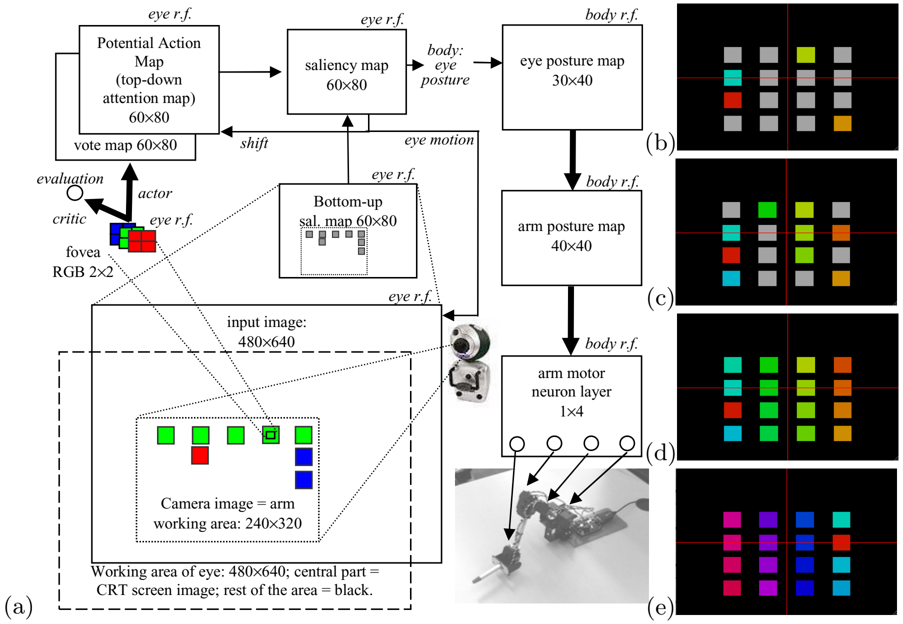


**Fig. 2.** (a) Eye-arm control architecture.(b-e) Scenes used in the experiment. (b) initial configuration with 0.9 chances of cue covered. (c) some covers removed. (d) map without the covers. (e) another example of map without the covers.

The rationale of this is that the informative role of an action should depend on the task and the environment. In the experiments, reaching actions can change the environment and uncover useful information to accomplish the task.

### 4.1 Experimental Setup

We used a task inspired from a card game for children. Two different families of 4x4 grid environments were used to train the agent: (a) in the first family the target is randomly positioned and all the other edges of the grid are occupied by cues, each of which has a precise spatial relationship with the target which is constant for all the environments in the family (see figure 2.d,e); (b) the second family is like the previous one but the cues are randomly hidden by grey *cover* (with each cue having 0.1 chance to be free since the beginning, see figure 2.b,c). In both families the target is never covered. When the agent touches a cover with the arm, the underlying cue is revealed. In both families of environments the agent can obtain reward only by touching the target. This will also start a new trial.

To use the nomenclature developed in section 2 is necessary to transform the second family of environments to a POMDP with one state for any possible configuration of the covers and for each position of the gaze and of the target. In this POMDP every state where the agent is not gazing to the target will have

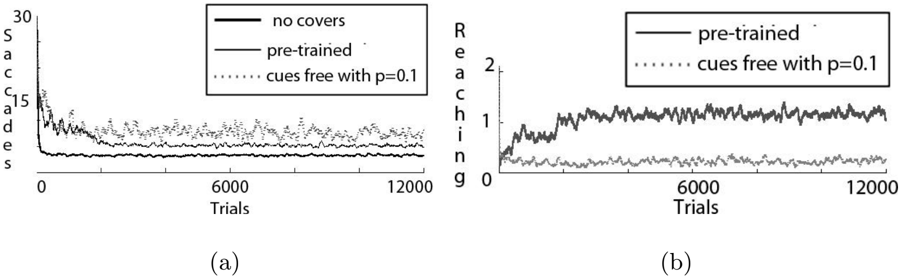


**Fig. 3.** (a) Average number of saccades per trial (b)Average number of reachings to covers per trial

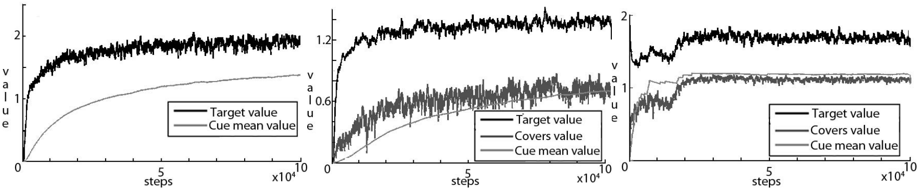
the same value in the underlying MDP because it is always possible to reach the goal state with the same number of eye and arm actions. So any gazing action from a state where cues or covers are gazed to a non targeted position is non-intrusive. Touching a cue or a cover is always non-intrusive because the target is uncovered from the beginning, thus in no condition removing a cover is necessary to obtain reward. Regarding informativeness, removing a cover is always informative since it gives access to a state that is unambiguous for the agent. Consequently, reaching a cue *is* an epistemic action in this context.

Three runs of 100,000 steps were executed in the two different environments. For each policy learnt with the clean map another run was executed with the corresponding environment with random covers. The data obtained in three runs for each condition were quite similar so only one run for condition is analysed in the following section.

### 4.2 Results

Figure 3.a shows the evolution of the average number of saccades per trial in the map task with agents fulfilling three different conditions: (a) learning with all the cues uncovered; (b) learning with most of the cues covered; (c) adapting from condition *a* to condition *b*. The final average number of saccades per trial for condition *a* is 3.2, for condition *b* is 8.0, and for condition *c* is 5.2. Thus, the agent in *c*, re-trained after having discovered the value of the cues, faces a simpler task then than the agent in *b* and so develops a better performance in the environment with the covered cues .

Figure 3.b shows the evolution of the average number of reaching actions on covers per trial (thus only conditions *b* and *c*). The final average number of reaching actions on covers for condition *b* is 0.3 and for condition *c* is 1.2. The agent in *c* uncovers a cue or more in most trials. It thus learns to execute strictly epistemic actions. Considering that a reaching action requires about 2-3 saccades to the same spot to be triggered, the exploration policy of the agent in *c* compared to that of the agent in *b* is more efficient than what might be supposed on the basis of the simple ratio between the number of saccades. While the agent

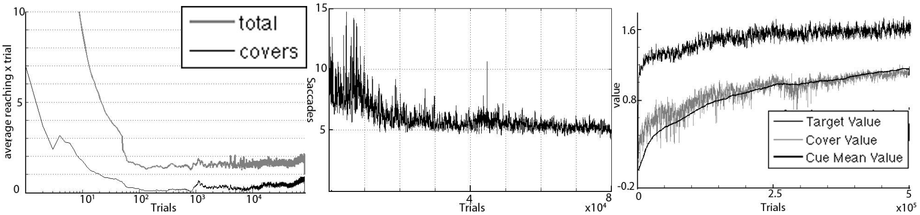


(a)agent a (b)agent b (c)agent c

**Fig. 4.** Values evolution for the various condition of training. In figure a,b and c the average value of the cue was plotted with a grey line.

in *b* and *c* is architecturally identical and operates in the same environment, it acquires very different behaviours due to the different training history.

Figures 4.a,b,c show the evolution of the value of the objects in the three conditions. The cues are represented by the evolution of the average value because the epistemic reaching action value depends on the average value of the cue that can be uncovered. The comparison of the three graphs shows that in condition *a* the values of the cues are learnt faster than in condition *b*. This means that initially in condition *b* the agent cannot learn to uncover the cues because it still does not know how to use the information it gets access to. Instead, looking away from the covers (which prevents reaching and uncovering them) can be useful because it can randomly lead to the target. On the other hand, when the agent knows the use of the cues, uncovering them is easily learnt if the agent has not inhibited this behaviour, as shown by Figure 4.c. This is a quite general main result of this research: for a model-free agent, the possibility of learning to use epistemic actions is strongly dependent on knowing how to use the information they deliver. Otherwise the epistemic actions can be inhibited and not explored/exploited anymore even when the agent later acquires the capacity to use the information they deliver. Probably, this might be ameliorated by using a mechanisms like internal simulation (like DYNA [27]). Even if in the condition *b*



(a)reaching (b)saccades (c)values

**Fig. 5.** Long run trial. (a) Evolution of average reaching per trial. (b) Evolution of average saccades. (c) Evolution of values of objects.

reported above the agent seemed to have converged to a stable performance level, a longer run (500,000 steps) with 10% of cues uncovered (same as condition *b*) was executed to test if it was able to learn to execute the epistemic actions with additional learning. The results are shown in figure 5. The final average number of saccades per trial is 5.1, similar to that of the agent in *c*. The final number of reaching actions on covers is 0.7, which means that in most of the trial the agent executes an epistemic action. However the agent takes about 50,000 trials to learn this behavior while the agent pre-trained on the clean map (c) develops the behavior in less than 2,000 trials (to which we must add the trials spent in the pre-training, that are about 12,000).

An experiment with an even longer training was executed with all the cues covered. In this condition even after 750,000 steps the agent was not able to develop the epistemic reaching action. In this condition a simple scanning procedure is really easy to learn for the agent, e.g., moving the gaze to the adjacent left element each time a cover is foveated. Instead, discovering the real underlying structure of the environment results in a complex policy involving a different action for each possible cue. This policy is particularly difficult to learn also because initially all the cues are covered, so moving from an uncovered cue to another position usually brings the agent to another cover.

## 5 Conclusions

Cognitive science research describes epistemic actions as aiming to change the internal state of the agent to (i) acquire new information from the environment, (ii) facilitate information processing, and (iii) acquire knowledge for better future processing and execution. The distinctive characteristic of epistemic actions is that they are executed by a bounded agent as a means to overcome its intrinsic perceptual, computational, and expertise limits [12, 13].

The work presented in this paper is focused on epistemic actions used to acquire information, named *external epistemic actions*. In psychology these actions have been named *specific exploration actions* [28]. The epistemic actions used to facilitate information processing can instead be named *internal epistemic actions*. A typical example from literature is rotating Tetris game blocks to facilitate visual matching instead of internally simulating their rotation [12]. Another is the use of sensorimotor strategies for discrimination instead of complex internal processing, e.g. scale and rotation invariance [20]. The third and last kind of epistemic actions are named *curiosity epistemic actions*: these are executed by the agent to increase its knowledge of the environment [28, 29].

We provided a formal definition of external epistemic actions for the POMDP framework, together with the concept of MPD-reducible-POMDP. This definition is dependent only on agent perceptual system and environment structure, so it can be applied without knowing the internal structure of the agent. Then we discussed several theoretical issues affecting a simple model-free agent using reinforcement learning in POMDP with epistemic actions, showing that even having MDP-reducible-POMDP is not a sufficient condition to permit learning of the optimal policy. We also showed the issues coming from (a) the initial ignorance of the use of the information that is accessed through an epistemic action and (b) the presence of suboptimal policies which are more informationally parsimonious and easy to learn. These policies tend to visit a limited set of perceptual states and show that the agent representation of the task, ignoring some of the hidden states, does not match with the actual environment. This is an important issue to consider when modelling organisms' behaviour using reinforcement learning. In this regards, it would be interesting to further study the preference for informationally parsimonious policies using a principled formal approach based on information theory like the one proposed in [30].

These issues were finally illustrated through a robotic experiment. Using different training procedures we showed the importance of knowing how to use the information acquired through an epistemic actions to learn the latter ones. Using an higher degree of partial observability resulted in the use of suboptimal strategies. In this respect, the results showed that it is possible to facilitate the acquisition of epistemic actions using shaping techniques [31]. The experiments also showed that the architecture proposed here is able to merge learnt epistemic actions with ontic actions in a smooth way in several conditions. On the contrary, most robotic architectures have a separate information acquisition phase followed by an action execution phase, thus limiting their adaptation capabilities.

Previous studies [20, 32] showed that reactive agents whose structure is developed with evolutionary algorithms can produce efficient behaviours even in partially observable environments. The results presented here extend these findings by showing that reactive agents can, in some conditions, learn through direct interaction with the environment how to incorporate epistemic actions in their policies, and this gives substantial advantages when adapting to complex environments with partial observability and a structure unknown at design time.

Another interesting further study can be to consider principles like "free-energy" minimization [33, 34]. When applied to the reduction of the uncertainty on the quantities related to the agent structure, e.g. the perceptual state of the agent, these might also result in the reduction of uncertainty on the environment hidden variables, similarly to epistemic actions.

## References

1. Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S.: Learning the value of information in an uncertain world. Nat. Neurosci. 10(9), 1214–1221 (2007)
2. Kepecs, A., Uchida, N., Zariwala, H.A., Mainen, Z.F.: Neural correlates, computation and behavioural impact of decision confidence. Nature 455(7210), 227–231 (2008)
3. Pezzulo, G., Rigoli, F., Chersi, F.: The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. Front Psychol. 4, 92 (2013)
4. Roy, N., Thrun, S.: Coastal navigation with mobile robots. In: Advances in Neural Information Processing Systems, vol. 12 (2000)
5. Cassandra, A., Kaelbling, L., Kurien, J.: Acting under uncertainty: discrete bayesian models for mobile-robotnavigation. In: Proc. of IROS 1996 (1996)

6. Kwok, C., Fox, D.: Reinforcement learning for sensing strategies. In: Proc. of IROS 2004 (2004)
7. Hsiao, K., Kaelbling, L., Lozano-Perez, T.: Task-driven tactile exploration. In: Proc. of Robotics: Science and Systems (RSS) (2010)
8. Lepora, N., Martinez, U., Prescott, T.: Active touch for robust perception under position uncertainty. In: IEEE Proceedings of ICRA (2013)
9. Sullivan, J., Mitchinson, B., Pearson, M.J., Evans, M., Lepora, N.F., Fox, C.W., Melhuish, C., Prescott, T.J.: Tactile discrimination using active whisker sensors. IEEE Sensors Journal 12(2), 350–362 (2012)
10. Moore, R.: 9 a formal theory of knowledge and action. In: Hobbs, J., Moore, R. (eds.) Formal Theories of the Commonsense World. Intellect Books (1985)
11. Herzig, A., Lang, J., Marquis, P.: Action representation and partially observable planning in epistemic logic. In: Proc. of IJCAI 2003 (2003)
12. Kirsh, D., Maglio, P.: On distinguishing epistemic from pragmatic action. Cognitive Science 18(4), 513–549 (1994)
13. Kirsh, D.: Thinking with external representations. AI & Society (February 2010)
14. Cassandra, A.R.: Exact and Approximate Algorithms for Partially Observable Markov Decision Processes. PhD thesis, Brown University (1998)
15. Melo, F.S., Ribeiro, I.M.: Transition entropy in partially observable markov decision processes. In: Proc. of the 9th IAS, pp. 282–289 (2006)
16. Denzler, J., Brown, C.: Information theoretic sensor data selection for active object recognition and state estimation. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(2), 145–157 (2002)
17. Whitehead, S., Lin, L.: Reinforcement learning of non-markov decision processes. Artificial Intelligence 73(1-2), 271–306 (1995)
18. Vlassis, N., Toussaint, M.: Model-free reinforcement learning as mixture learning. In: Proc. of the 26th Ann. Int. Conf. on Machine Learning, pp. 1081–1088. ACM (2009)
19. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
20. Nolfi, S.: Power and the limits of reactive agents. Neurocomputing 42(1-4), 119–145 (2002)
21. Aberdeen, D., Baxter, J.: Scalable internal-state policy-gradient methods for pomdps. In: Proc. of Int. Conf. Machine Learning, pp. 3–10 (2002)
22. Whitehead, S.D., Ballard, D.H.: Learning to perceive and act by trial and error. Machine Learning 7(1), 45–83 (1991)
23. Koenig, S., Simmons, R.G.: The effect of representation and knowledge on goal-directed exploration with reinforcement-learning algorithms. Mach. Learn. (1996)
24. Ognibene, D.: Ecological Adaptive Perception from a Neuro-Robotic perspective: theory, architecture and experiments. PhD thesis, University of Genoa (May 2009)
25. Ognibene, D., Pezzulo, G., Baldassarre, G.: Learning to look in different environments: An active-vision model which learns and readapts visual routines. In: Proc. of the 11th Conf. on Simulation of Adaptive Behaviour (2010)
26. Balkenius, C.: Attention, habituation and conditioning: Toward a computational model. Cognitive Science Quarterly 1(2), 171–204 (2000)
27. Sutton, R.S.: Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proc. ICML, pp. 216–224 (1990)
28. Berlyne: Curiosity and exploration. Science 153(3731), 9–96 (1966)
29. Baldassarre, G., Mirolli, M.: Intrinsically Motivated Learning in Natural and Artificial Systems. Springer, Berlin (2013)

30. Tishby, N., Polani, D.: Information theory of decisions and actions. In: Perception-Action Cycle, pp. 601–636. Springer (2011)
31. Ng, A.Y., Harada, D., Russell, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Proc.of the ICML, pp. 278–287 (1999)
32. Beer, R.D.: The dynamics of active categorical perception in an evolved model agent. Adapt. Behav. 11, 209–243 (2003)
33. Friston, K., Adams, R.A., Perrinet, L., Breakspear, M.: Perceptions as hypotheses: saccades as experiments. Frontiers in Psychology 3 (2012)
34. Ortega, P.A., Braun, D.A.: Thermodynamics as a theory of decision-making with information-processing costs. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 469(2153) (2013)