# Combining artificial curiosity and tutor guidance for environment exploration

Pierre Fournier, Olivier Sigaud, Mohamed Chetouani

# Combining artificial curiosity and tutor guidance for environment exploration

Pierre Fournier[1], Olivier Sigaud[1] and Mohamed Chetouani[1]

***Abstract*— In a new environment, an artificial agent should explore autonomously and exploit tutoring signals from human caregivers. While these two mechanisms have mainly been studied in isolation, we show in this paper that a carefully designed combination of both performs better than each separately. To this end, we propose an autonomous agent whose actions result from a user-defined weighted combination of two drives: a tendency for gaze-following behaviors in presence of a tutor, and a novelty-based intrinsic curiosity. They are both incorporated in a model-based reinforcement learning framework through reward shaping. The agent is evaluated on a discretized pick-and-place task in order to explore the effects of various combinations of both drives. Results show how a properly tuned combination leads to a faster and more consistent discovery of the task than using each drive in isolation. Additionally, experiments in a reward-free version of the environment indicate that combining curiosity and gaze-following behaviors is a promising path for real-life exploration in artificial agents.**

## I. Introduction

Mental development for a situated agent includes the capacity to actively discover how to achieve various tasks in an unknown environment. Young children demonstrate very early the inclination to explore their surroundings, trying to interact with objects within their reach. In artificial systems, these behaviors are the subject of Artificial Curiosity. Taking inspiration from trial-and-error exploration, some of the existing work is built on the Reinforcement Learning (RL) framework [1]. Such solutions, called intrinsically motivated reinforcement learning (IMRL) [2], rely on attractive salient events [3] or search to maximize metrics of their environment model [4]. A limitation of the RL framework for autonomous environment exploration is the notion of external reward. Indeed, such rewards are most commonly human-engineered task-specific functions [5] that cannot be defined in the essentially task-independent exploration problem, nor redefined for every new environment.

When in presence of a caregiver, infants are also able to incorporate social stimuli into their exploration behavior (following a tutor gaze, copying gestures, ...), long before the acquisition of an effective language-like high-level communication channel. In this context, social signals can favor task exploration and guide it towards the expectations of a tutor. Such ability to exploit skills and knowledge from humans in artificial systems is the subject of Interactive Machine Learning (IML) [6]. IML algorithms can be organized along a scale between focusing entirely on human guidance [7] versus relying only on autonomous exploration [8]. Between those extremes, some combination of human guidance and autonomous exploration have proven efficient, for instance to learn object affordances [9].

The present work pushes the latter approach further by proposing a generic RL system where the balance between the motivation to explore and the motivation to interact is easily tunable, and shows that the right proportion of each leads to increased performances on task-discovery and task execution in a new environment. To this end, we propose an agent whose actions respond to a user-defined weighted average of two incentives: artificial curiosity on one side, and a tendency to follow the gaze of a tutor on the other side.

The artificial curiosity component is based on the intrinsically motivated RL algorithm of TEXPLORE-VANIR [4]. The gaze-following component, which is considered essential to learn to interact [10], [11], is implemented through the addition of gaze direction in the TEXPLORE-VANIR agent and the use of a new reward shaping mechanism that entices the agent to match its gaze with the tutor's.

After presenting TEXPLORE-VANIR in Section II, our method and the simulated task are described in Section III. Results are presented in Section IV. First, we show that with a simple gaze policy for the tutor, the right weights for both curiosity and gaze-following behaviors lead to improved performances for task discovery and execution. Furthermore, we show that our agent can consistently complete the task in reward-free environments. Finally, Section V discusses the originality of this work in comparison with the existing literature, gives potential directions for future work and concludes.

## II. Background: TEXPLORE-VANIR

The TEXPLORE-VANIR algorithm is an improvement of the TEXPLORE model-based reinforcement learning framework [12] that includes additional intrinsic motivations.

### A. TEXPLORE

The TEXPLORE algorithm follows a typical model-based RL scheme, where the environment is modeled as a factored Markov Decision Process (MDP). An MDP consists of a set of states $S$, a set of actions $A$, a reward function $R(s,a)$ and a transition function $T(s,a,s')$. The agent receives the reward $R(s,a)$ upon taking action $a$ in state $s$ and ends up in state $s'$ with probability $P(s'|s,a) = T(s,a,s')$. The agent seeks to determine the policy $\pi^\star : s \mapsto a$ that maximizes the expected

*This work was not supported by any organization

[1] The authors are with: Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005 Paris, France Contact: pierre.fournier@isir.upmc.fr +33 (0) 1 44 27 88 53

discounted total reward over the agent lifetime. Defining the state-value function $Q^{\pi}(s,a)$ as an estimate of the expected future reward obtainable from $(s,a)$ following policy $\pi$, $Q^{\star} = Q^{\pi^{\star}}$ solves the Bellman equation [1] and

$$\pi^{\star}(s) = \arg\max_{a} Q^{\star}(s,a). \tag{1}$$

Being model-based, the TEXPLORE agent learns models of $R$ and $T$ from experience to simulate multiple courses of actions and iteratively refine its Q-table of $Q(s,a)$.

In TEXPLORE, model learning is seen as a self-supervised problem, with the current factored state $s = (s_1, \ldots, s_n)$ and action $a$ as inputs, and the next state $s'$ and the obtained reward $r$ as output. The factorization property of the model enables learning separate predictions for each state features $s_i$ (and for the reward) and recombine them thereafter into a full predicted state. This is only valid under the assumption that all features transition independently. In TEXPLORE, each separate feature predictor is a random forest comprising five univariate C4.5 decision trees, trained on different subsets of the experiences. Tabular models of the reward and transitions are re-built from the newly trained predictors when outdated, and are queried by the planner during simulations with the UCT algorithm [13]. These simulations provide action-reward-state sequence $a_t, r_{t+1}, s_{t+1}, \ldots, r_{t+M}, s_{t+M}$, where $M$ is the maximum simulation depth. All along the simulated sequence, Q-value updates for state-action pair $(s_{t'}, a_{t'})$ write:

$$Q(s_{t'},a_{t'}) \xleftarrow{\alpha_{UCT}} (1-\lambda) \sum_{i=1}^{M+t-t'} \lambda^{i-1} \times \left[ \left( \sum_{k=1}^{i} \gamma^{k-1} R(s_{t'+k-1}, a_{t'+k-1}) \right) + \gamma^{i} \max_{a'} Q(s_{t'+i}, a') \right]. \tag{2}$$

The notation $x \xleftarrow{\alpha} y$ is short for $x \leftarrow (1-\alpha)x + \alpha y$, $\lambda$ is the eligibility traces parameter and $\gamma$ is the discount factor. $\alpha_{UCT}$ is a learning rate specific to the UCT algorithm [12]. In TEXPLORE, $R(s,a) = r_{pred}(s,a)$ as rewards only comprise the reward predictions by the reward model model given input state and action.

### B. Novelty reward in TEXPLORE-VANIR

In TEXPLORE-VANIR, the reward $R$ in (2) is enriched by two intrinsic motivations favoring 1) high-uncertainty areas of the environment and 2) high-novelty areas of the state-action space. The present work only uses the second one based on novelty.

For a given state-action pair $(s,a) = (s_1, \ldots, s_n, a)$, this additional NOVELTY-REWARD $N(s,a,V)$ is based on the normalized distance between $(s,a)$ and the set $V$ of known state-action pairs kept up to date. It writes:

$$N(s,a,V) = \arg\min_{(s_V,a_V)\in V} \|(s,a) - (s_V,a_V)\|_1. \tag{3}$$

## III. METHODS

We augment the TEXPLORE-VANIR algorithm in two ways: 1) the agent's action achievements are now conditioned to their coordination with its gaze, and 2) a second reward shaping mechanism is added, in order to favor states in which the agent gaze follows the tutor's.

**Algorithm 1** TEXPLORE-VANIR with guidance by gaze-following

1: **Input:** An actor, a tutor and an environment
2: Initialize $Q(s,a) = 0$, $\forall s,a$
3: Environment model M $\leftarrow$ empty model
4: Starting state $s \leftarrow s_0$, known states $V \leftarrow \varnothing$
5: $\pi_{tutor} \leftarrow$ predefined policy, tutor state $\sigma \leftarrow \sigma_0$
6: **loop**
7: $\quad a \leftarrow \arg\max_{a'} Q(s,a')$
8: $\quad$ Actor takes action $a$, observes $r, s'$
9: $\quad$ Tutor updates gaze $\sigma$ following $\pi_{tutor}(s')$ ▷
10: $\quad$ TRAINPREDICTORS$(\langle s,a,s',r\rangle, M)$
11: $\quad V \leftarrow V \cup (s,a)$
12: $\quad \sigma^{obs} \leftarrow \sigma$ ▷
13: $\quad$ **for all** state $s$, action $a$ **do**
14: $\quad\quad T(s,a), R(s,a) \leftarrow$ UPDATEENVMODEL$(s,a,M)$
15: $\quad\quad R(s,a) \mathrel{+}= N(s,a,V)$
16: $\quad\quad R(s,a) \mathrel{+}= J(s,\sigma^{obs})$ ▷
17: $\quad$ **end for**
18: $\quad$ UCTPLANNING$(Q,T,R)$
19: $\quad s \leftarrow s'$
20: **end loop**

### A. Gaze following motivation

To fulfill the action-attention coordination constraint, we augment the agent state $s$ with gaze information so that we now have $s = (s^{env}, s^{gaze})$ where $s^{env}$ is comprised of pure environment observations from the agent while $s^{gaze}$ is its gaze position. The list of possible actions for the agent in state $s = (s^{env}, s^{gaze})$ is now conditioned to $s^{gaze}$. The next section provides details on these conditions for our specific experimental framework.

To induce gaze-following behaviors, a reward $J$ is given if the agent's gaze matches the tutor's: for a given state $s = (s^{env}, s^{gaze})$, and a given observation by the agent of the current tutor gaze $\sigma^{obs}$, we write:

$$J(s,\sigma^{obs}) = \delta(s^{gaze}, \sigma^{obs}) \tag{4}$$

where $\delta$ is the Kronecker symbol.

### B. Algorithm

The final algorithm structure is shown in Alg. 1, where additional steps with respect to TEXPLORE-VANIR are highlighted. Lines 7-12 comprise building experience that includes the tutor's reactions and training the feature predictors from this experience. Lines 13-17 perform the tabular models updates from the predictors. The reward computations thus include incentives for novelty and gaze-following behaviors by shaping $R$ in (2) with (3) and (4), writing:

$$\begin{aligned} R(s_{t'+k-1}, a_{t'+k-1}) = {} & r_{pred}(s_{t'+k-1}, a_{t'+k-1}) \\ & + \nu N(s_{t'+k-1}, a_{t'+k-1}, V) \\ & + \mu J(s_{t'+k-1}, \sigma^{obs}_{t}), \end{aligned} \tag{5}$$

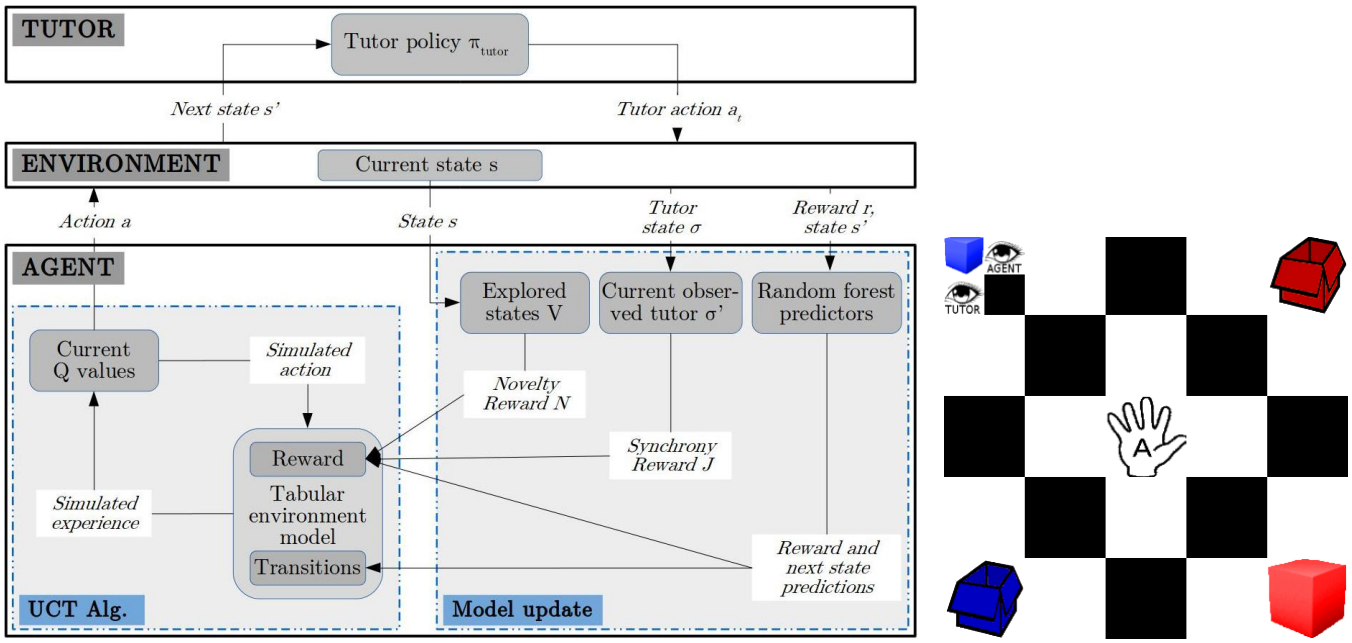


Fig. 1. **Left**: the agent model is an extension of the classical model-based reinforcement learning scheme, where the tutor behavior is explicitly taken into account to build experience. The agent exploits the accumulated experience to train predictors, from which the tabular environment model is built. Novelty and gaze-based motivations modulate the tabular reward model to favor specific state-action pairs. The agent obtains the best action by computing Q values from simulations based on the environment model. **Right**: The environment is a 5x5 grid with two sources of red and blue blocks (the cubes) and a box of each color. The agent is defined by its position (the hand) and its actions need to be coordinated with its gaze (its eye). The tutor only exists through its gaze, and he looks where it is best for the agent to also look at.

where $\nu$ and $\mu$ are tunable parameters that determine the importance granted to novelty and gaze-following behaviors respectively. It is important to note that $R(s_{t'}, a_{t'})$ relies on $\sigma_t^{obs}$ for all $t'$ considered in (2): the tutor gaze is not part of the agent state and thus no prediction is made about its future. Thus updates at simulated step $t'$ use the value $\sigma^{obs}$ at step $t$.

Finally line 18 corresponds to the UCT-PLANNING algorithm where the Q-table is updated following (2) and (5). Figure 1 (left) shows the full model-based learning framework with explicit use of the tutor gaze signals to modulate the tabular reward model.

### C. Experiments

The algorithm is evaluated on a virtual pick-and-place task in a 5x5 grid world environment, shown in Fig. 1, right. Two infinite sources of blocks are located in two corners of the grid, while two boxes are located in the remaining corners. The agent must go to a block, pick it up, carry it to a box and place it inside. The agent (the hand in Fig. 1) is positioned at the center of the grid at the start of an experiment. The agent must repeatedly put blocks in a box. The positions of the blocks and boxes do not change over time or over trials. The agent state consists of 13 different features: its $x_a$ and $y_a$ coordinates in the room, whether it carries a block, the $x_g$ and $y_g$ coordinates of its gaze, and the $x_i$ and $y_i$ coordinates of all four objects in the grid.

Ten actions are available: the agent can move one step in each cardinal direction, look at all four objects, pick a block and put one in a box. Apart from looking at an object which is always possible, an action realization is conditioned to the environment and to the agent gaze. A move cannot be made against a wall, and the move must be in the direction of the agent gaze. To be successful, the pick action (resp. put action) requires the agent to be located at a block position (resp. a box position), while holding nothing (resp. holding a block); its gaze must be directed towards the block (resp. the box). All actions are deterministic and when one is not successful, the agent state remains unchanged.

The tutor's policy in this environment is fixed. Each time the agent ends up holding nothing during exploration, the tutor picks a source of blocks, choosing randomly. Then the tutor keeps looking at the source until the agent takes a block from it. When the agent picks a block, irrespective if it is the one the tutor looked at, the tutor chooses a box at random and looks at it until the agent has placed the block in it. This policy is an oversimplification of natural gaze mechanisms, based on object-directed visual fixation and does not aim at reproducing natural interaction behaviors.

The parameters $\lambda$, $\gamma$ and $M$ are those of TEXPLORE-VANIR and remain constant in all experiments: $\lambda = 0.1$, $\gamma = 0.9$ and $M = 100$. Q-values are initialized uniformly at random around zero. All experiments consists of 30 trials starting in the exact same conditions (parameters, starting position, etc.). Each trial comprises 800 learning steps. In addition to the accumulated reward, we store for each action $a$ taken in state $s$ the proportions of each of the three rewards in $Q(s,a)$.

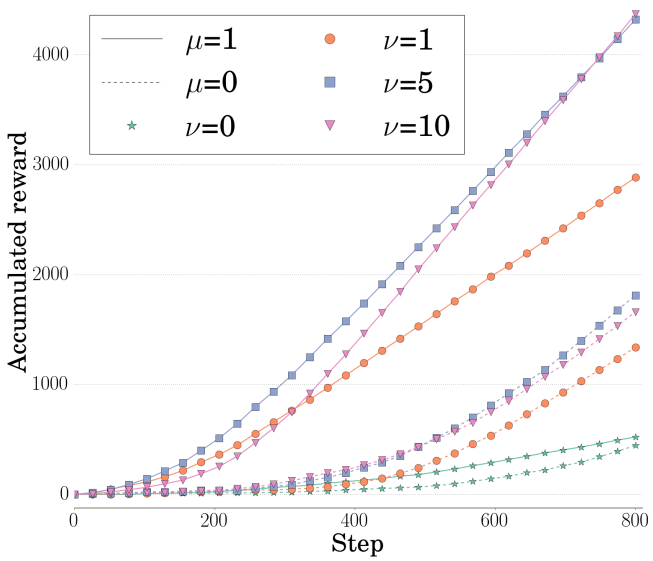


Fig. 2. Accumulated reward versus number of steps taken, averaged over 30 trials, with and without a tutor ($\mu = 0$ or $\mu = 1$) for different parameters $\nu$. The incentive to follow the gaze of the tutor clearly leads to better policies. Performances are best for $\mu = 1$ and $\nu = 10$.

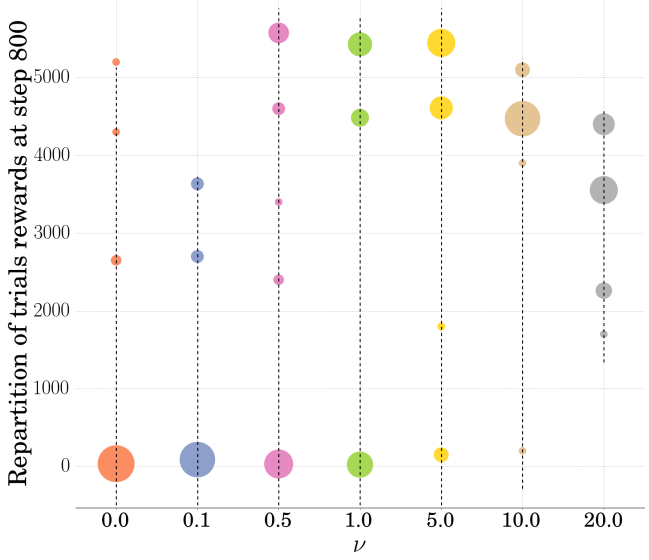


Fig. 3. Distributions of the accumulated reward at step 800 over the 30 trials for different values of $\nu$, each corresponding to a vertical line, for a fixed $\mu = 1$. On each vertical, the 30 trial results are binned in the six $[0,1000]$, $[1000,2000]$, ..., $[5000,6000]$ intervals. Each bin is then displayed as a circle at height equal to the average value of the bin and of size proportionate to the number of results in the bin. The best result in term of both accumulated reward and consistency over trials is obtained for $\nu = 10$: few low reward results and many high reward results.

## IV. Results

The present model was first designed to measure the impact of the relation between the weight of the incentive for gaze-following and that of curiosity on task discovery in a new environment. In a second phase we evaluate our full motivation system in a reward free version of the environment.

### A. Task-oriented exploration

First, the contribution of the tutor during reinforcement learning is examined. Performances of the agent on the pick-and-place task with and without taking the tutor gaze into account ($\mu = 1$ and $\mu = 0$) are compared. As we are mainly interested in guiding exploration, we focus on the first few hundreds iterations. Fig. 2 plots the accumulated rewards versus the number of steps taken by the agent, averaged over 30 trials. The figure shows results for different intrinsic motivation weights: $\mu = 0$ and $\mu = 1$ on one side, $\nu = 0$, $\nu = 1$, $\nu = 5$ and $\nu = 10$ on the other. Experiments where the agent takes the tutor gaze into account appear more efficient at discovering the task than their counterparts with intrinsic motivation only. Differences between results for $\nu = 5$ and $\nu = 10$, be it with or without a tutor, are not meaningful on this plot and require further investigation.

Figure 2 shows different performance gains for a constant appeal to joint attention ($\mu = 1$) depending on the importance granted to curiosity through $\nu$. This suggests a coupling between intrinsic motivation and gaze-following, which we analyze in Fig. 3. To this end, we make $\nu$ vary between 0 and 20 with a fixed $\mu = 1$ and focus on the end reward accumulated at step 800 only. For each value of $\nu$, Fig. 3 shows on a vertical line the distribution of results obtained over the 30 trials. To account for trials giving identical results, these 30 values are binned in intervals $[0,1000]$, $[1000,2000]$, ..., $[5000,6000]$ and each bin is shown as a circle centered at the average value of the bin, and of size proportionate to the number of elements in the bin. For instance, with no curiosity at all ($\nu = 0$, on the left), most trials have not discovered the task nor obtained any reward, hence the large dot on the no-reward horizontal axis.

A significant proportion of the trials for $\nu \leq 10$ still obtain few to no reward at all, as illustrated by a remaining circle close to zero; on the other hand, the majority of other trials reach high end results as shown by large dots for high intervals. Between these two behaviors lie very few samples. These results indicate that either the agent discovers the task early enough and then exploits its discovery to reach a high final accumulated reward, or it remains stuck in a form of inefficient exploration and receives no reward at all. The absence of intermediate end-result (few values between 1000 and 4000 for $0.5 \leq \nu \leq 10$ in particular) is a consequence of this alternative. If we evaluate the chosen parameters based on consistency over trials and on the actual 800 step accumulated reward, performances are best for $\nu = 10$ and $\mu = 1$. The 10/1 ratio between both motivations approximately compensates the fact that the bonus obtained from novelty is always rather small compared to that from the gaze (the right side of (3) is rarely close to 1).

The observed behavior is easily interpreted in the light of the role played by each reward mechanism during exploration. To measure these roles, we use the proportions of each reward/motivation mechanism inside the Q-value corresponding to the chosen action at each iteration, which

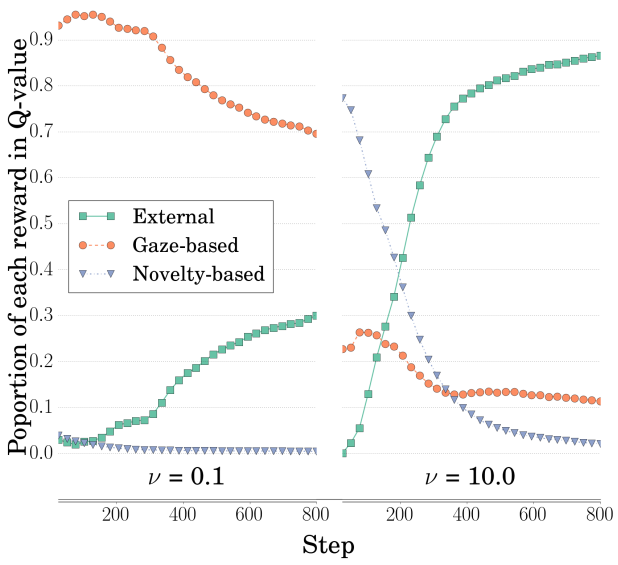


Fig. 4. Importance of each reward mechanism in action-selection during the first 800 steps, for the two combinations $[\mu = 1, \nu = 0.1]$ (left) and $[\mu = 1, \nu = 10]$ (right). The evolution of the proportion of each member of (5) in the Q-value defining the next action is displayed. Successful combinations of motivations (right) enable novelty to play the main role at the very beginning and give way to environment rewards when they are discovered, while the tutor guidance impacts the agent with a continuous moderate intensity. An insufficient curiosity leads to following the tutor gaze only (left).

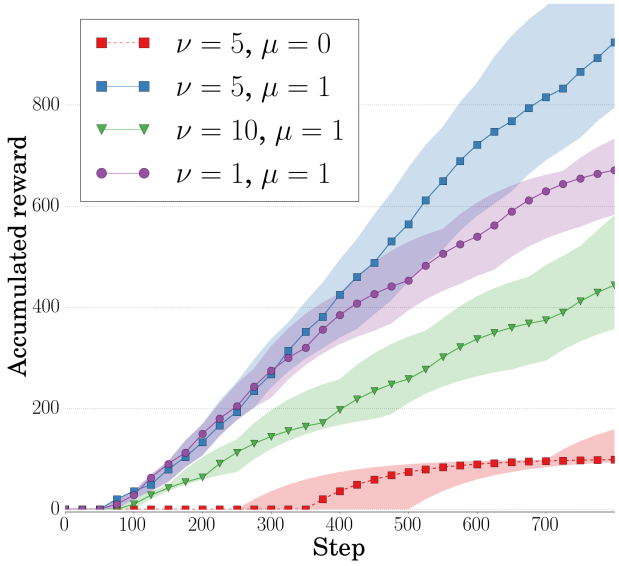


Fig. 5. Evolution of the reward accumulated for putting blocks in boxes over 800 steps (median and 25-75 interquartile range), for different combinations of gaze-based and curiosity-based motivations. Proper combinations of curiosity and interaction drives lead to the task being achieved regularly without using dedicated external rewards.

we store all along the trials. They indicate how much each reward/motivation mechanism is responsible for the action of the agent. The evolution of these proportions, shown in Fig. 4 on the left for two very different behaviors (weak versus strong curiosity), indicates that low performances on the left of Fig. 3 correspond to the agent only looking where the tutor looks, without searching to discover the environment and with growing but weak interest for the task. The decrease in performance observed for high values of $\nu$ stems from the opposite behavior where the agent's goal is the complete discovery of the environment, independently of the task or the tutor. The benefit of this strategy is that it almost guarantees random task execution and avoids no-reward trials. A successful trade-off between curiosity, joint attention and external reward exploitation is reached for $\nu = 10$ (Fig. 4, right): novelty-based motivation appears critical at the beginning of learning and then gives way to external reward signal discovered, while the tutor acts as a permanent discrete guidance. This combination leads to a more effective task-oriented exploration of the environment than any of the drives taken separately.

### B. Reward-free environments

Now the reward obtained from the environment is only used for evaluation purposes and learning only relies on the two last terms of (5). The agent is only driven by curiosity and the tutor gaze. Fig. 5 shows the impact of the attention-based guidance on the accumulated reward corresponding to the amount of blocks put inside boxes over time. With curiosity only, the agent achieves the task by chance on rare occasions over the 800 steps. By contrast, in presence of a tutor, it discovers the task earlier and more importantly, also achieves the task expected by the tutor with regularity. This ability is obtained through the combination of the two reward mechanisms and the attention-action coordination constraint: schematically the agent attention is first driven to the parts of the state space indicated by the tutor; because of the action-attention coordination constraints, the agent itself is then more likely to actually end up in those states the tutor deems useful; finally, once in those potentially rewarding states, the curiosity mechanism ensures that the agent will find the right action, among those it has not tried yet.

The plot also shows that the binary behavior described in the previous section has been reduced. Indeed, the 25-75 interquartile ranges drawn for each experiment over the 30 corresponding trials indicate that there is far less uncertainty in the end-results without external rewards. Also, the coupling between parameters $\nu$ and $\mu$ still exists without reward: $[\nu = 5, \mu = 1]$ performs significantly better than both $[\nu = 10, \mu = 1]$ and $[\nu = 1, \mu = 1]$ as shown by the non-overlapping interquartile ranges. A comparison with Fig. 2 shows that learning without external reward logically remains less efficient than with it. This is coherent with the earlier explanation of task successes: with an external reward obtained when the goal is reached, the agent can propagate it backward in its sequence of actions, so that it knows how to act in each state, which is not possible without reward propagation.

## V. Conclusion

In this paper, we presented a generic extension of the RL framework to combine autonomous exploration out of

curiosity, and guidance from a tutor based on gaze-following. Contrary to a number of works in IML, such as RL with on line human-defined rewards [14], our solution does not consider interaction as a secondary tool to fulfill a primary objective. Instead, interaction is seen as a goal in itself, and is favored by a dedicated reward mechanism. Also, interaction takes a bottom-up approach and relies on a low-level gaze-based social cue as in other works [15]. Previous studies following this line of thought have faced difficulties in interpreting such social cues, as it adds complexity to the problem [16], [17], [18]. Our solution tackles the issue by converting gaze direction into an exploitable guidance signal through reward shaping in the RL framework.

Results show that adding a tutor gaze direction as guidance to a curious RL agent leads to improved exploratory abilities, provided that curiosity and motivation for gaze-following are combined in correct proportions. When such a successful balance is reached, the tutor gaze acts as a constant and moderate push towards profitable states, while curiosity appears decisive at the beginning of exploration, and fades as it goes on. With this behavior, the agent performs better at task-discovery and execution than other combinations of curiosity and guidance by tutoring signals, or without guidance at all.

As the few models that have tackled the issue of autonomous reward-free environment discovery, our curiosity mechanism mainly relies on the agent evaluating the accuracy of its internal model of the environment [19], [20]. This evaluation then serves as a basis for directing exploratory behavior. Our work demonstrates that the balanced exploitation of a guidance signal is also an effective solution to speed up task discovery and regular execution in such reward-free environments.

Future work should further elaborate the gaze following mechanism which is hitherto oversimplified: the tutor gaze direction is given whereas it should be detected. This causes the agent to follow its tutor too easily instead of learning gaze following [21]. In the end, we would like to allow a tutor to teach an agent a specific task without using any external reward signal. Steps in this direction will result from integrating automatic goal-discovery [22] and tutor intentions inference [23]. Both would enable subgoal-based rewards and competence-based motivations [24] that would speed up exploration and improve guidance efficiency. Such work should be performed in larger state spaces so as to evaluate the scalability of the presented algorithm and its performances with very sparse rewards.

## REFERENCES


[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[2] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? A typology of computational approaches," *Frontiers in Neurorobotics*, vol. 1, 2009.

[3] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, pp. 1281–1288, 2004.

[4] T. Hester and P. Stone, "Intrinsically motivated model learning for a developing curious agent," in *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pp. 1–6, IEEE, 2012.

[5] D. Dewey, "Reinforcement learning and the reward engineering principle," in *2014 AAAI Spring Symposium Series*, 2014.

[6] E. Senft, S. Lemaignan, P. E. Baxter, and T. Belpaeme, "Leveraging Human Inputs in Interactive Machine Learning for Human Robot Interaction," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 281–282, ACM, 2017.

[7] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli, "A decision-theoretic model of assistance," *Journal of Artificial Intelligence Research*, vol. 50, no. 1, pp. 71–104, 2014.

[8] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. L. Thomaz, "Policy Shaping: Integrating Human Feedback with Reinforcement Learning," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 2625–2633, Curran Associates, Inc., 2013.

[9] V. Chu, T. Fitzgerald, and A. L. Thomaz, "Learning Object Affordances by Leveraging the Combination of Human-Guidance and Self-Exploration," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, (Piscataway, NJ, USA), pp. 221–228, IEEE Press, 2016.

[10] I. Nomikou, G. Leonardi, K. J. Rohlfing, and J. Rczaszek-Leonardi, "Constructing Interaction: The Development of Gaze Dynamics," *Infant and Child Development*, vol. 25, pp. 277–295, May 2016.

[11] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, no. 2, pp. 135–169, 2006.

[12] T. Hester and P. Stone, "TEXPLORE: real-time sample-efficient reinforcement learning for robots," *Machine learning*, vol. 90, no. 3, pp. 385–429, 2013.

[13] L. Kocsis and C. Szepesvari, "Bandit-Based Monte-Carlo planning," in *European conference on machine learning*, pp. 282–293, Springer, 2006.

[14] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in *Proceedings of the fifth international conference on Knowledge capture*, pp. 9–16, ACM, 2009.

[15] F. Broz, H. Kose-Bagci, C. L. Nehaniv, and K. Dautenhahn, "Learning behavior for a social interaction game with a childlike humanoid robot," in *Social Learning in Interactive Scenarios Workshop, Humanoids*, 2009.

[16] A. Najar, O. Sigaud, and M. Chetouani, "Social-Task Learning for HRI," in *International Conference on Social Robotics*, pp. 472–481, Springer, 2015.

[17] M. Lopes, T. Cederborg, and P. Y. Oudeyer, "Simultaneous acquisition of task and feedback models," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2, pp. 1–7, Aug. 2011.

[18] J. Grizou, M. Lopes, and P.-Y. Oudeyer, "Robot Learning Simultaneously a Task and How to Interpret Human Instructions," in *Joint IEEE International Conference on Development and Learning an on Epigenetic Robotics (ICDL-EpiRob)*, (Osaka, Japan), Aug. 2013.

[19] P.-Y. Oudeyer, F. Kaplan, V. V. Hafner, and A. Whyte, "The playground experiment: Task-independent development of a curious robot," in *Proceedings of the AAAI Spring Symposium on Developmental Robotics*, pp. 42–47, Stanford, California, 2005.

[20] D. Y. Little and F. T. Sommer, "Learning and exploration in action-perception loops," *Frontiers in Neural Circuits*, vol. 7, Mar. 2013.

[21] E. Carlson and J. Triesch, "A computational model of the emergence of gaze following," *Progress in Neural Processing*, vol. 15, pp. 105–114, 2004.

[22] S. Forestier, Y. Mollard, D. Caselli, and P.-Y. Oudeyer, "Autonomous exploration, active learning and human guidance with open-source Poppy humanoid robot platform and Explauto library," Dec. 2016.

[23] A. O. Diaconescu, C. Mathys, L. A. E. Weber, J. Daunizeau, L. Kasper, E. I. Lomakina, E. Fehr, and K. E. Stephan, "Inferring on the Intentions of Others by Hierarchical Bayesian Learning," *PLOS Computational Biology*, vol. 10, p. e1003810, Sept. 2014.

[24] S. Forestier and P. Y. Oudeyer, "Towards hierarchical curiosity-driven exploration of sensorimotor models," in *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 234–235, Aug. 2015.