This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

Selective Perception As a Mechanism To Adapt Agents To The Environment: An Evolutionary Approach

Mirza Ramicic and Andrea Bonarini

***Abstract*—Rapid advancement of machine learning makes it possible to consider large amounts of data to learn from. Learning agents may get data ranging on real intervals directly from the environment they interact with, in a process usually time-expensive. To improve learning and manage these data, approximated models and memory mechanisms are adopted. In most of the implementations of reinforcement learning facing this type of data, approximation is obtained by neural networks and the process of drawing information from data is mediated by a short-term memory that stores the previous experiences for additional re-learning, to speed-up the learning process, mimicking what is done by people.**

In this work, we are proposing a novel computational approach able to selectively filter the information, or cognitive load, for the agent's short-term memory, thus emulating the attention mechanism characteristic of human perception. We devised an evolutionary model of agent's perception to adapt the *attention filter* present in the proposed architecture to the actual environment faced by the agent, by selecting the experiences that most likely influence in a positive way its learning characteristics. This approach can evolve a filter able to provide an optimal cognitive load of the experiences entering in the agent's short-term memory of a limited capacity. The evolved sampling dynamics can also lead to the emergence of intrinsically motivated curiosity.

***Index Terms*—Machine Learning, Intelligent Agents, Cognition, Genetic Algorithms, Artificial Neural Networks, Attention, Perception, Short-term memory.**

I. Introduction

The evolutionary development of a human brain was followed by its ability to perceive and elaborate an increasing amount of external stimuli from its immediate environment. However, no matter how much the brain processing power increased, its ability to process the whole amount of incoming sensory data is still limited. At the current stage of development our brains receive through sensors millions of bits of information each second, but we are able to consciously process only about 126 bits over that time interval [1]. The existence of this bottleneck brought to the development of another mechanism, capable of focusing on the subset of information actually useful for the cognitive process. This cognitive filter that protects us from the sensory overload is called *attention* [1], [2]. A major role in this filter is played by a system named *working memory*, which acts as a buffer between our perception and its conscious processing [3]. The *working memory* consists of a temporary memory storage along with specialized mechanisms for replaying its contents. It also relies on the central, or executive, attention that is in charge of regulating its active contents therefore providing a specific memory-related context for the higher order cognitive processing [4]. By now, a growing number of psychological and neuro-scientific studies have confirmed that the *working memory* is selective when it comes to storing stimuli: sensory stimuli that are important to the goal show enhanced activity, while the other, irrelevant stimuli are suppressed [5], [6], [7].

In contemporary machine learning, similarly, the ability to process large continuous state spaces grew with the introduction of function approximators such as artificial neural networks [8]. Recent implementations, such as [9], [10], include a memory buffer called *replay memory* that is functionally similar to the human *working memory*: it selectively stores the experiences, or transitions, in order to replay and re-learn from them off-line, therefore reducing the amount of data that should be acquired by expensive processes, and ensuring at the same time a more stable training of the approximator needed to manage continuous variables. Later approaches that dealt with the mechanism of *replay memory* were interested in improving the speed of learning by focusing the attention on specific transitions that are more valuable to the learning process, using criteria such as temporal difference error [11], received reinforcement [12], and information potential of the state [13].

Usually, in machine learning, the agent prefers unexpected experiences as they are more likely to "surprise" the predictor and feed the learning process in order to further reduce the uncertainty about the environment [14], [15]. The preference for novel experiences is also found in the developing human brain, as the babies not only prefer, but are driven to focus attention on situations that include novelty and surprise [16], [17], [18] in order to acquire much needed stimulation. These discoveries have led to a concept called *intrinsic motivation* in machine learning, which deals with the problems of motivating novelty seeking for its own sake [19], [20], [21], [22].

In this work, we introduce a computational model of *attention* that includes a mechanism for selectively storing experiences from the agent's perceptive stream into its *replay memory*, therefore providing a goal-related context buffer from which the past experiences can be sampled for re-learning. Furthermore, we explore how it is possible for an artificial agent to evolve this attention mechanism over generations so to make it learning more efficiently by selectively focusing on experiences more valuable than others in a given environment. After multiple trials over 100 generations, we have found that the focusing mechanism can learn to be more selective; this is consistent with the concept of attention, and the *replay*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

memory contents get permeated with transitions that are high in the values of curiosity indicative parameter *informational gain* [14]. Evolutionary motivated, higher level goals superseded the primary reinforcement-based ones, making the agent to adopt a new long-term strategy of intrinsic curiosity, while exhibiting a behavioral change of being more brave or aggressive towards acquiring new sources of stimuli in the given environment. However, this attitude may be more or less effective, depending on the characteristics of the specific environment, thus calling for a mechanism to adapt curiosity to them. We are proposing an evolutionary approach implementing such a mechanism.

II. Approximation

When facing high dimensional state spaces it is highly impractical for a reinforcement learning system to keep the estimates of the Q (quality) value for each combination of state and action, and it is even impossible if we are dealing with continuous state representations. In this case, the best option is to approximate the estimation of the optimal value, $Q^*(s,a)$ using a function approximator such as an *artificial neural network*, or ANN. A function approximation makes it possible to predict a Q value for each of the possible actions available to the agent by providing an agent's current state in input to the ANN. After each time step we can compute the expected Q value using Bellman equation and compare it to the estimate that the function approximator provides as its output $Q(s_t, a_t; \Theta) \approx Q^*(s_t, a_t)$ by providing any state s_t in input. The difference between the previous estimate of the approximator and the expectation is the TD error, and it is possible to back propagate it through the ANN in order to update the current approximation of $Q^*(s,a)$.
The backpropagation is performed by updating the weights of the ANN approximator Θ by performing a gradient descent on the loss $L_i(\Theta_i)$ according to Equation 1:

$$\nabla_{\Theta_i} L_i(\Theta_i) = (y_i - Q(s,a;\Theta_i))\, \nabla_{\Theta_i} Q(s,a;\Theta_i), \quad (1)$$

where $y_i = r + \gamma \max_{a'} Q(s', a'; \Theta_{i-1})$ is actually the Bellman equation defining the target value.

III. Related Work

Probably the earliest successful use of the past experiences to support the direct update reinforcement learning algorithms such as Q-learning can be seen in Dyna-Q [23].
The Dyna-Q algorithm re-uses direct experiences in order to build and constantly update a model of the environment which predictions are then used to generate simulated virtual experiences. Using Dyna-Q an agent is able to learn not only from direct experiences, but also from virtually generated ones. This allows an agent to be more effective in updating its value functions with a limited amount of actual experience.

A. Intrinsic Motivation

In more recent developments we can see the use of a competence-based, intrinsic motivation in supporting the generation and selection of agent's goals. Active goal exploration *SAGG-RIAC* strategy [24] is able to use intrinsic motivation in order to select at each time step a goal that maximizes the progress of competence in reaching the goal in previous experiences. *GRAIL* framework [25] takes a step further and makes use of intrinsic motivation in order to also generate a set of goals that an agent is able to focus on while learning. This process relies on the perceived differences between the current and previous sensed state, which is similar to the concept of informational gain *IG* used by the approach we are proposing to represent agent's intrinsic curiosity. Florensa et. al [26] recently presented an improvement over *SAGG-RIAC* that is able to explore the agent's goal space more efficiently along with a novel goal generation mechanism. In this approach a Generative Adversarial Network or *GAN* is able to produce goals within an optimal dynamic difficulty given the agent's current progress.

B. Artificial attention as a behavior inducing mechanism

The idea that a selective focusing of the memories that enter in the short-term memory analogue of *replay memory* can behaviorally influence the artificial learning agents was explored in [27], by considering biases that may come from personality traits or attitudes of the agents towards exploration. In this approach, a computational model of main personality trait axis including introversion-extroversion dichotomy was developed. The model was based only on changing the dynamics of the attention span of the *replay memory*, which was different between introverted and extroverted individuals, as the latter exhibited a broader attention span [4]. The two types of agents were tested in different configurations of the environment characterized by different amount of reinforcement. *Normal* environment provided an equal distribution of positive and negative reinforcement and represented the baseline for the experiment. *Hostile* environment provided more negative reinforcement, while the *benevolent* one provided more positive reinforcement. Curiosity-driven, extroverted agents performed better in a *benevolent* type of environment while the cautious, introverted ones managed to learn better in the *hostile* environment.
Attention-based working memory approach was used also in [13], which proposed a selective focusing of the experiences based on their information potential or Shannon entropy of the perceived state space. Although the aim of this approach was primarily to increase the learning performance of the agents, compared to an uniform sampling baseline the entropy-based sampling also seemed to have induced an intrinsically motivated exploration that became an important part of their tactics to increase the overall performance.
Persiani et al. [28] introduces an approach that also uses the replay memory structure in order to improve cognition. The algorithm is able to actively learn to select the most appropriate chunks of the agent's experience to be stored in the replay memory buffer based on maximizing the expected future reward.

C. Evolutionary Adaptive Approaches

In [29] basic emotions such as fear were evolved as motivational drives involved in adaptation of learning agents to their

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

immediate environment. At each generation, a new population of virtual agents was tested, each of them evolving a neural network that mapped its input, which included time, good and bad sensation neurons, and visual perception, to its output that was used to focus on the visual stimuli and select the agent's actions. Over time, the selection of agents based on elitism w.r.t. the ability to adapt to the environment gave rise to a specific drive of being cautious or fearful as a survival strategy. Another evolutionary perspective is presented in [30] in which the reward functions of the agents are evolved by taking in consideration their fitness. This forms an idea of *optimal reward function* that builds upon the basic reward function in order to maximize the expected fitness over distribution of environments. The presented experiments support the notion of emergence of intrinsic reward for specific actions such as playing and manipulating objects in their immediate environment that are not meeting any primary need of the agent. [31] introduced a combination of evolution and machine learning allowing the agents to intrinsically evolve a basic reinforcer for atomic building-block skills in the *childhood* learning phase, which are later used in the *adulthood* phase.

IV. Model Architecture and Learning Algorithm

In this section, the structure of the learning model we are proposing is presented as its two main parts: evaluation and evolution.

The first part, shown in Figure 1, represents the main evaluation reinforcement model, where the proposed *attention focus block* or *AFB* plays a primary role. *AFB* is a filtering mechanism which receives in input a raw stream of the experiences that are perceived by the agent, and selects the ones that will be stored in the replay memory for re-learning purposes. It does so using its main component: ANN function approximator (f), which receives in input the characterizing parameters of the transition between states and produces as output a crisp decision about whether to sample the transition into the replay memory, or to discard it.

The architecture of the AFB neural network approximator (f) consists of three layers: three input nodes connected to a fully connected hidden layer of four nodes, in turn connected to two softmax nodes to produce the final classification. This ANN is able to approximate the three parameters of the experience, respectively given in input as TD error, entropy of the starting state s_t, and *informational gain*, to a probability of belonging to a class "sample" or "do not sample".

The main part of the learning algorithm is the *learning loop*, represented as the (b) section in Figure 1, where the agent actually takes an action a that brings it from state s_t to the next state s_{t+1}, which is also providing an immediate reinforcement r_t. The actual learning part of the loop is supported by a main function approximator ANN shown in (d) block of Figure 1, which performs a backpropagation update at each learning step in order to provide a better approximation of the Q values of the state-action pairs. This ANN takes a multi-dimensional state on its input and provides the estimated Q values for each possible action available for the agent as its output layer. Since the target value for $Q(s, a)$ is given by the Bellman equation, it is possible to calculate it, taking into account the immediate reinforcement and the discounted Q value of the next state, and to compare it with the current estimate of the function approximator ANN in order to figure out how wrong it was with respect to what obtained by the last transition. This difference, also known as *TD error* provides enough information to update the new estimation of $Q(s, a)$. A *backpropagation* is performed on the approximator with the state s_t on the input layer and the gradient on the a_t output is set to *TD error* while all other gradients on the action outputs are set to 0. After the update, the transition is actuated so that s_t becomes s_{t+1} and the *loop* restarts.

The *learning loop* process provides the raw sequential experience stream (marked as (c) in Figure 1 from which it is possible to sample some of the experiences into a buffer structure called *replay memory* (a), which stores the experience transitions from which the agent can selectively learn. In this proposal, this process is mediated by the *attention focus block*, including another ANN function approximator (f) able to predict whether or not it is the case to sample the specific transition in the *replay memory*, depending on its properties provided in input. The properties forwarded through the ANN encompass the predictive power, given by its *TD error*, along with its information potential factors such as the Shannon's entropy of the state s_t and the *information gain* potential of the transition given by Kullback-Leibler difference between state s_{t+1} and s_t.

The evolution of the proposed *attention focus block* is obtained by a *genetic algorithm*, where the members of the population encode the information about the weights of the ANN approximator (f), used to control the sampling dynamics of the *attention focus block* structure. The starting point for the algorithm is a random population, and each configuration of parameters for an *attention focus block* dives learning for an agent operating in the environment. Each agent performs a learning process mediated by its *genetically* altered *attention* in an environment that provides both positive and negative reinforcement. Over a given number of learning steps, we may observe that agents learn to gravitate towards positive and to avoid negative reinforcement sources. The total reward collected at the end of a trial represents the agent's score, which is indicative of its adaptability to the specific environment.

After the evaluation phase, the agents are ranked by the total reinforcement they received, which is taken as their respective fitness. The *mating* phase starts by selecting the individuals with the higher fitness that are going to be the base for the next population. The selected genotypes undergo the genetic operations of crossover and mutation. In crossover, two *parent* genotypes randomly combine their genetic information to produce an offspring. Then, mutation randomly modifies the genetic material. The resulting genotypes produced by these two operations join to form a new evolved generation from which the process of evaluation can start again.

V. Experimental Setup

The proposed approach was tested in different environments. Here are presented the results obtained in two environments, derived from some of the environments used in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

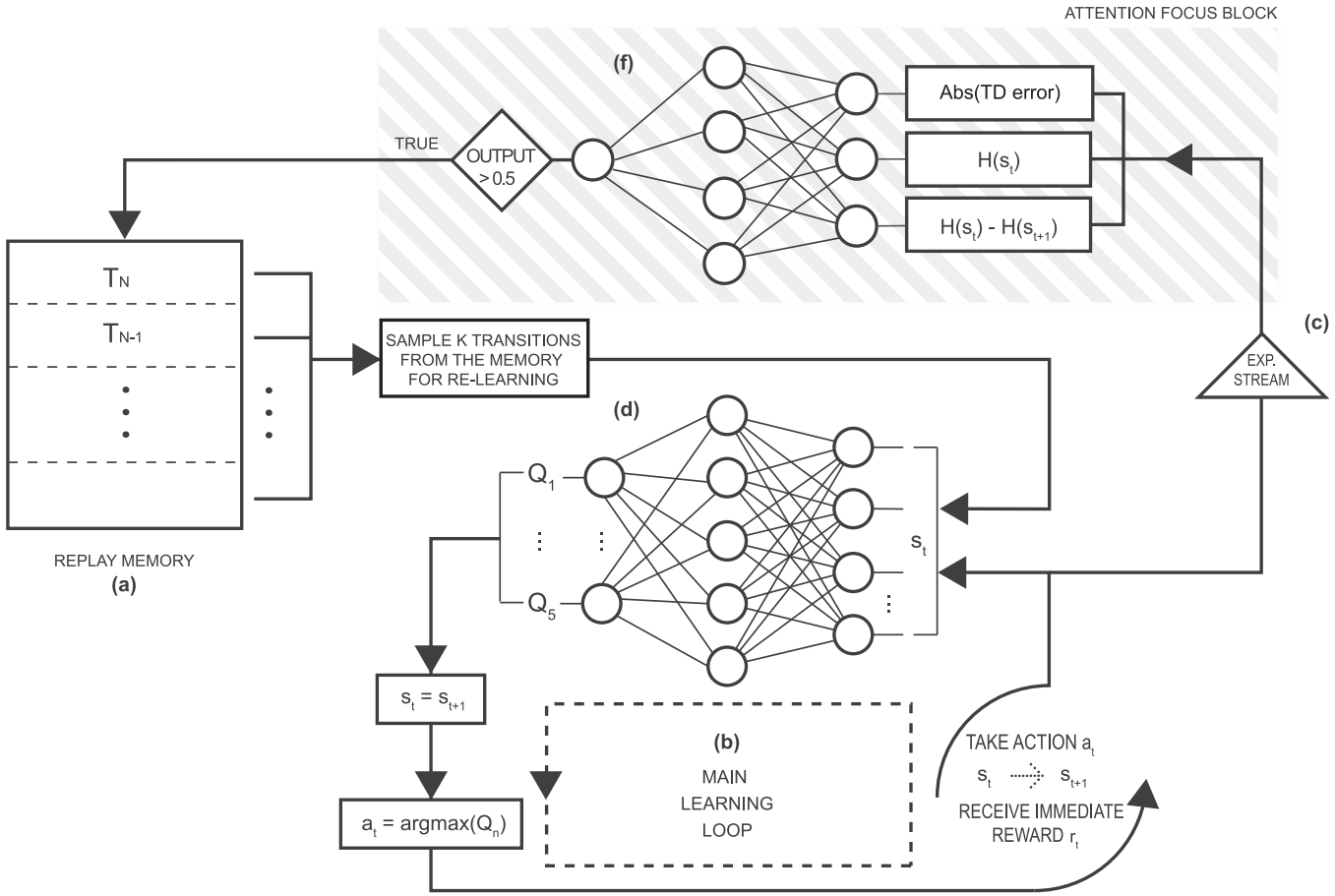


Fig. 1. General learning model architecture including *attention focus block*: (a) Replay memory; (b) Main learning loop; (d) A block implementing main Q-value function approximator neural network ; (c) Raw stream of the experiences; (f) ANN function approximator as a part of *attention focus block*

literature, namely *Waterworld* [32], and *LunarLander-v2* from *OpenAI Gym* framework [33]. In both cases we have a complex and continuous state spaces which can support their diversity.

A. Waterworld Environment

In the Waterworld environment the agents operate in an environment inspired by the *Waterworld* setting showcased in [32]. It consists of moving food pieces that are instantiated with random speed and direction and are capable of bouncing off the walls surrounding a closed space. The food pieces come in two dispositions: good, which provides a positive reinforcement of $+1$, and bad, which results in a negative reinforcement of -1, upon contact with an agent. The environment always contains an equal amount of good and bad food: once the food has been consumed, a new food source is randomly generated in order to keep the distribution constant. The agent's goal consists of maximizing its expected reward in the long run, so it has to learn to consume as much of good food pieces while avoiding the bad ones. The agent is able to move while accumulating momentum or stay still by taking five possible actions: left,right,up,down and stay. Its perception of the environment is implemented as 30 equally distributed directional sensors, each capable of perceiving five continuous variables: distances to the perceived object, being it good food, bad food or walls along with additional two variables for velocity components of the detected object in x and y. This, along with the perception of two additional variables for the agent's own velocity (x and y components), gives a quite high dimensional state-space consisting of 152 continuous variables.

The Q value function approximator ANN (marked as (d) in Figure 1 is implemented as three layers: the input layer consists of 152 nodes fully connected to an inner layer of 100 nodes, that is in turn connected with an output layer of 5 nodes (the possible actions) and trained using a learning rate $\alpha = 0.005$. It is able to approximate the 152 dimensions of state on the input to the Q values of 5 actions available to the agent. In the genetic algorithm adapting the approximator ANN, the genes were modified with a heuristically determined probability of 0.25 and the modification was implemented as the addition of a number between -0.1 and 0.1 to the respective parameter value.

The evaluation trial lasted 160,000 steps. Reinforcement learning expectation is computed with a discount rate $\gamma = 0.9$, and $\epsilon-$greedy policy is used with the starting $\epsilon = 0.2$, then adjusted to 0.1 at the mid-point of the trial, after 80.000 steps for a better convergence towards the end. Replay memory buffer capacity was set to 3000 experience transitions. A

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

population of 4 agents was evolved and after each evaluation the best performing two were selected for crossover and mutation. The new population was made from a multiple mating of the two best performing agents.

B. Lunar Lander Envrionment

The second set of experiments was performed in a more realistic setup, such as *LunarLander-v2* from *OpenAI Gym* framework [33]. The goal is to land a craft in a designated landing place indicated by two flags while countering the gravity pull using three thrusters: main, left and right orientations. While *Waterworld* is representative of tasks with continuous variables and sparse, random reinforcement, *Lunar Lander* is representative of rocket trajectory optimization which is a classic topic in area of optimal control featuring a more dispersed and constant reinforcement feedback. An episode concludes when a lander crashes or comes to a rest, in which case it receives additional -100 or +100 of reinforcement respectively. Additional reinforcement is provided, inversely proportional to the craft distance from the landing area and deviation from zero speed; it comes in the range of +100 to +140. Firing main thruster results in a -0.3 reinforcement while each leg contact with a ground is rewarded by +10. The craft has an unlimited amount of fuel at its disposal and can also land outside the designated area.

State space is 8-dimensional and consists of 4 continuous variables sensing the x position of the craft, its y position relative to the land area, craft's angle and its angular velocity along with 2 boolean variables indicating a land contact for each of the craft's leg. Four discrete actions are available to the craft: do nothing, fire main engine, fire left orientation engine, fire right orientation engine.

Reinforcement learning parameters were set to be the same as in previous batch of experiments along with an adjustment of ϵ. Elitism was implemented allowing two best scoring agents to propagate their genotypes unchanged into the next generation while the other 8 phenotypes of the next generation were generated by crossover of genotypes selected with a probability proportional to their respective scores. Mutation rate was also 0.25 and again performed by adding a number between -0.1 and $+0.1$ to the parameter value and attention filter block ANN architecture is the same as in previous batches with an exception of a hidden layer containing 6 neurons which slightly increased the variety of genotypes.

VI. Experimental Results

The experiments performed on both environments, *Waterworld* and *Lunar Lander*, were compared using three settings. A genetic algorithm evolutionary implementation of the proposed *attention focus block* sampling, or *GA-AFBS*, was compared with a non-evolutionary case *R-AFBS* of generations consisting of randomly selected weight parameters, and a baseline *NO-AFBS* in which agents used no *AFBS* filtering and sampled every experienced transition into the replay memory buffer.

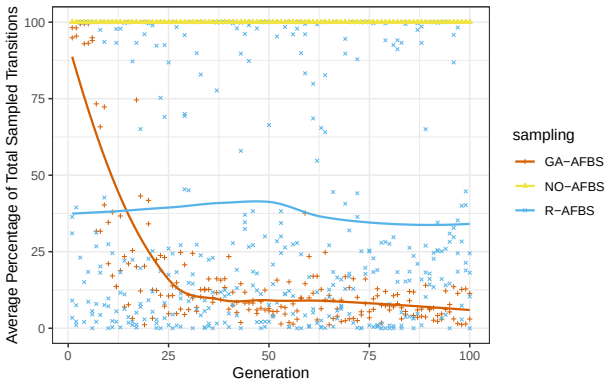


Fig. 2. Average total sampling percentage in the *Waterworld* environment of the genetic algorithm supported evolution of the *attention focus block* (*GA-AFBS*) compared with a non-evolutionary sampling implemented as a random attention filter neural network (*R-AFBS*) and a baseline approach without any cognitive filter (*NO-AFBS*), over the first 100 generations of 6 trials (solid lines) and the respective variance (gray areas).

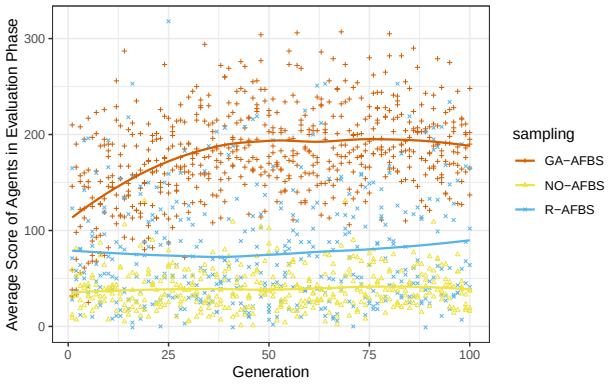


Fig. 3. Average fitness (total reinforcement) received in *Waterworld* environment by the proposed genetic algorithm supported evolution of the *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS*, and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

A. Waterworld Environment

A total of 6 trials were performed in the *Waterworld* environment, each of them evolved a 100 generations of *attention focus block* phenotypes evaluated by reinforcement learning phases for $160,000$ learning steps.

In Figure 2 we can observe the evolution of the number of experiences sampled by the *attention focus block* over 100 generations. Experimental data show that the *attention focus block* evolved in the direction of being more selective about the sampled experiences, from inefficiently taking almost 88% of the raw experience stream in the replay memory at the beginning, to a much more selective selection of 12% experiences at the 100th generation, which represents a great difference with respect to the random *R-AFBS* percentage which was constantly kept around 40%.

Figure 3 shows how the evolutionary model influenced the performance of the agents given by their total score, or total reinforcement received over the evaluation phase. An approach that used evolving phenotypes of *attention focus block* (*GA-*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

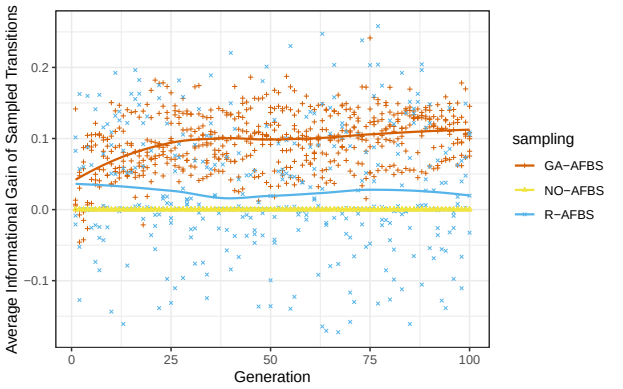


Fig. 4. Average values of *informational gain* parameter of experiences contained in *working memory* at the end of evaluation in *Waterworld* environment for the proposed genetic algorithm supported evolution of *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS* and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

AFBS) greatly outperformed the no filter one (*NO-AFBS*) by over 400% and a random parametrization (*R-AFBS*) by more than 200%, and came to a stable point in about 75 generations. Figure 4 shows the sampling preference of the approaches in terms of types of transitions as determined by average *informational gain* values of the sampled experiences in the *replay memory*. We can see that the genetically supported evolution of *GA-AFBS* evolved a high tendency to sample the experiences with positive values of the *informational gain* property in contrast with the no filter *NO-AFBS*, whose average accumulated to 0 as it preferred the experiences with both positive and negative values in the same proportion. The evolutionary approach settled to a 0.1 average informational gain which gave rise to a more curious agents than a random *R-AFBS* one, which was rather consistent with an average of 0.25 throughout the generations.

B. Lunar Lander Environment

Evaluation phase in *Lunar Lander* environment in *GA-AFBS* consisted of a generation of 10 agents competing with each other based on the average reward received over 60 consecutive learning episodes.

Changes in sampling preferences can be seen from Figure 5, where the proposed evolutionary approach *GA-AFBS* evolved again to be more restrictive to select experiences for replay memory. Although not as selective as *Waterworld*, in the *Lunar Lander* environment *GA-AFBS* resulted in a more conservative 25% sampling percentage which is a significant change compared to the expected 50% average sampling of the random network *R-AFBS*.

Figure 6 shows the amount of improvement that *GA-AFBS* brought to the more realistic *Lunar Lander* environment. We can see that *GA-AFBS* took about 12 generations to outperform the baseline *NO-AFBS* by 150% and provide a 125% increase over the random *R-AFBS*.

We can also notice that *GA-AFBS* evolved a preference for sampling experiences that manifest a higher value of the starting state entropy $H(s_t)$ outlined by the Figure 7 and a high preference for experiences with lower informational gain level IG which can be seen from Figure 8.

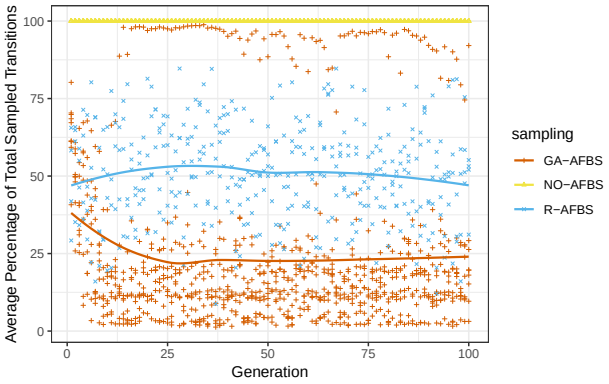


Fig. 5. Average total sampling percentage in *Lunar Lander* environment for the proposed genetic algorithm supported evolution of *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS* and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

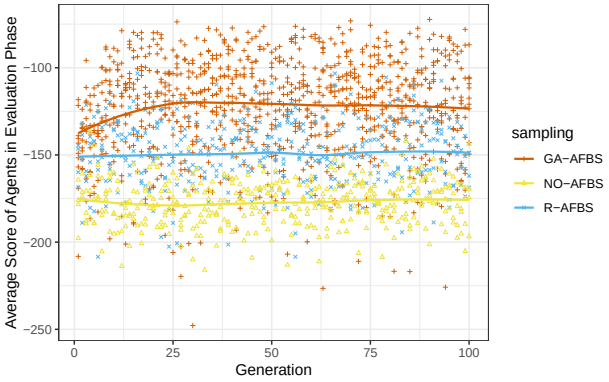


Fig. 6. Average fitness or total reinforcement received in *Lunar Lander* environment for the proposed genetic algorithm supported evolution of *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS* and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

VII. Discussion

A. Informational Gain parameter as a measure of Curiosity

Informational gain or *IG* parameter is defined as Kullback-Leibler difference or relative entropy between posterior state s_{t+1} and anterior one s_t, as summarized in Equation 2. It is especially important for a discussion about the emergence of intrinsically motivated evolved behavioral traits of agents in the *GA-AFBS* sampling method. It can provide an insight about the agent preference to move towards a state of higher informational content, which is indicative of intrinsic curiosity if it is positive, while, on the other end of the spectrum, negative values are indicating a more cautious move in which the agent is moving away from the state of high informational potential.

$$IG = H(s_{t+1}) - H(s_t) \tag{2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

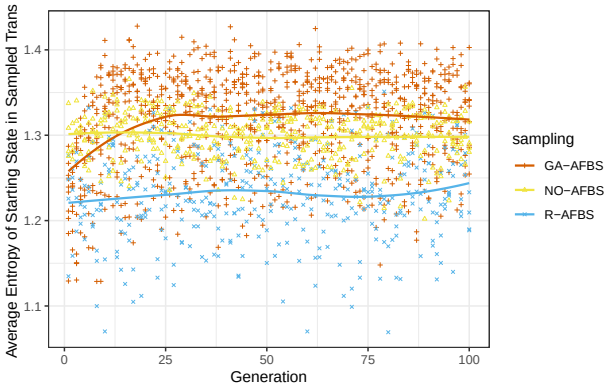


Fig. 7. Average values of *starting state entropy* parameter of experiences contained in *working memory* at the end of evaluation in *Lunar Lander* environment for the proposed genetic algorithm supported evolution of *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS* and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

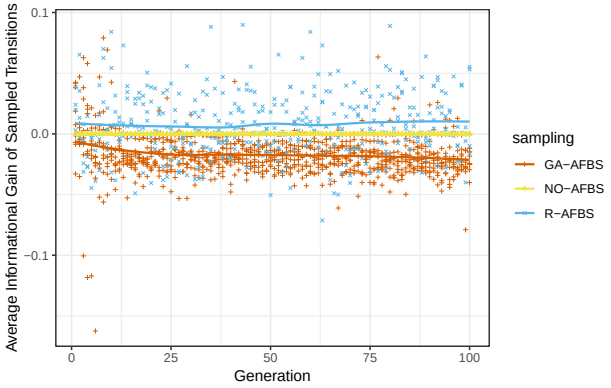
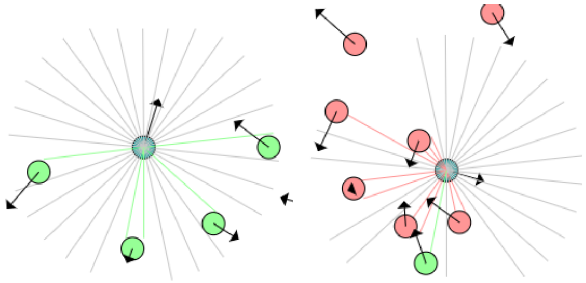


Fig. 8. Average values of *informational gain* parameter of experiences contained in *working memory* at the end of evaluation in *Lunar Lander* environment for the proposed genetic algorithm supported evolution of *attention focus block GA-AFBS* sampling, compared with a non-evolutionary sampling implemented as random attention filter neural network *R-AFBS* and a baseline approach without any cognitive filter *NO-AFBS* over first 100 generations of 6 trials.

This behavioral characteristics can be seen in Figure 9 and Figure 10, which compare the transitions with respectively low and high levels of *informational gain* parameter by showcasing the velocity vectors of both agent and food in a specific transition in the *Waterworld* environment. From Figure 9 we can see that the experiences that are low in *IG* tend to transition the agent to a "safer zone" by moving it away from the food, making it safer w.r.t. the risk of collecting bad food, but also making it less likely to consume good food. On the other end, a more curious agent will prefer the transitions that are shown in Figure 10, and that tend to push the agent to a more "interesting" state, where it is expected to collect new experience, as given by the positive value of the *IG* parameter.

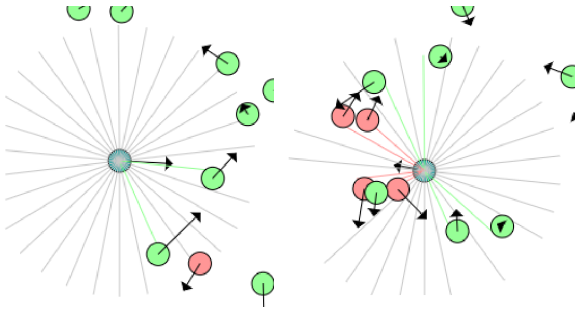
B. Implications

From Figure 2 and Figure 5, we can conclude that the evolutionary approach of *GA-AFBS* can reduce the cognitive load on the agent induced by a highly saturated, high-dimensional



(a) $IG = -0,1528$ (b) $IG = -0,2083$

Fig. 9. Transitions with low *Informational Gain* values in the *Waterworld* environment.



(a) $IG = 0,2872$ (b) $IG = 0,3064$

Fig. 10. Transitions with high *Informational Gain* values in the *Waterworld* environment.

state space, by selecting more interesting experiences that are stored for learning in the *replay memory*. Besides evolving an optimal cognitive load for each of the considered environments of 12% for the *Waterworld* and 25% for the more realistic *Lunar Lander*, this approach also improved the selection of experiences that are more valuable for machine learning as evident from Figure 3 and Figure 6, which show a significant improvement of the total reinforcement received in both environments. The fact that the different optimal sampling percentages were evolved in adaptation to the environments bring us to a conclusion that the different environments present a varying level of cognitive load for the learning agent. The more chaotic nature of the *Waterworld* gave rise to more interesting, information saturated transitions, which resulted in a more selective perception compared to the not so saturated one in *Lunar Lander* environment.

Some behavioral characteristics such as curiosity were intrinsically evolved to better adapt to the specific dynamics of the environment. From Figure 4 we can see that in the *Waterworld* environment curiosity or positive *IG* was evolved as an adaptation trait that arose from the need of an agent to be more engaged in the environment with a scarce reward and more focused in finding transitions that lead to situations expected to provide positive reinforcement.
Contrary to the scarce reinforcement feedback of the *Waterworld* environment, *Lunar Lander* provided a totally different and more dynamic reward mechanism which included constant adjustment of the reinforcement function based on the agent's

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2896306, IEEE Transactions on Cognitive and Developmental Systems

state. Confronted with the dynamics of the *Lunar Lander* environment, *GA-AFBS* evolved a trait of being cautious given its preference for the transitions with negative *IG* as shown in Figure 8. Also interesting to note is that the evolved perception mechanism in *Lunar Lander* displayed a preference for the transitions that have a higher entropy of the starting state, as can be see in Figure 7, which possibly contained more informational potential for learning, but at the same time cautiously preferred low entropy of the next state s_{t+1} given by the negative *IG* displayed in Figure 8.

From displayed results it is possible to notice that an evolved artificial perception using the proposed *GA-AFBS* algorithm was able to alter the behavioural characteristics of the agents by producing agents with specific traits or tactics that provide a better adaptation to a specific environment without the need to alter the reinforcement function.

VIII. Conclusion

As machine learning mechanisms evolve, we are now aware that along the advancement of the learning algorithms focused on how to learn from data received from the environment in a most efficient way, we also need to be concerned about the way those data are *perceived* in the first place. In spite of being still vastly unexplored, a good source of inspiration for new computational and evolutionary approaches of *perception* is a computational organ that is a product of a million of years of evolution: the human brain.

In this work we tried to exploit the insights received by the areas of psychology and neuroscience, which describe the higher order complex functions that our brain is using in order to make its perception more efficient. Since these functions were developed in humans by an evolutionary process of natural selection, it seemed that a similar process in a computational sense would also do the job. Although oversimplified compared to the human brain, the proposed approach could develop a filtering mechanism able to reduce the cognitive load and to induce an effective intrinsic behavior, therefore simulating the arising of the process of *perception* just by changing the dynamics of experience sampling. Furthermore, the genetically evolved perception mechanism was able to improve the learning performance by optimizing the way the agent collects and makes use of the information potentially provided by its environment. This brings us to the notion that the perception modeling in reinforcement learning can be seen as a more than a rigid, one-time, feature design, but can be implemented by a dynamic and state responsive mechanism that be, by itself, capable of eliciting behavioral adaptations to the environment characteristics during the learning process. In the future work, we will explore how to learn the best parameters for the *attention focus block* in order to further improve its selection capabilities.

References

[1] M. Csikszentmihalyi, "Flow: The psychology of optimal performance," *NY: Cambridge UniversityPress*, vol. 40, 1990.

[2] M. W. Eysenck, *Attention and Arousal, Cognition and Performance*. Springer, 1982, vol. 1, no. 1.

[3] A. D. Baddeley and G. Hitch, "Working memory," *Psychology of learning and motivation*, vol. 8, pp. 47–89, 1974.

[4] R. W. Engle, "Working memory capacity as executive attention," *Current directions in psychological science*, vol. 11, no. 1, pp. 19–23, 2002.

[5] J. W. de Fockert, G. Rees, C. D. Frith, and N. Lavie, "The role of working memory in visual selective attention," *Science*, vol. 291, no. 5509, pp. 1803–1806, 2001.

[6] P. E. Downing, "Interactions between visual working memory and selective attention," *Psychological science*, vol. 11, no. 6, pp. 467–473, 2000.

[7] A. Gazzaley and A. C. Nobre, "Top-down modulation: bridging selective attention and working memory," *Trends in cognitive sciences*, vol. 16, no. 2, pp. 129–135, 2012.

[8] L.-J. Lin, "Reinforcement learning for robots using neural networks," DTIC Document, Tech. Rep., 1993.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.

[12] J. Zhai, Q. Liu, Z. Zhang, S. Zhong, H. Zhu, P. Zhang, and C. Sun, "Deep q-learning with prioritized sampling," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 13–22.

[13] M. Ramicic and A. Bonarini, "Entropy-based prioritized sampling in deep q-learning," in *Image, Vision and Computing (ICIVC), 2017 2nd International Conference on*. IEEE, 2017, pp. 1068–1072.

[14] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in neurorobotics*, vol. 1, p. 6, 2009.

[15] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[16] R. W. White, "Motivation reconsidered: The concept of competence." *Psychological review*, vol. 66, no. 5, p. 297, 1959.

[17] D. E. Berlyne, "Conflict, arousal, and curiosity." 1960.

[18] E. J. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge," *Annual review of psychology*, vol. 39, no. 1, pp. 1–42, 1988.

[19] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2005, pp. 1281–1288.

[20] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007.

[21] P.-Y. Oudeyer and F. Kaplan, "How can we define intrinsic motivation?" in *Proceedings of the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Lund: LUCS, Brighton*. Lund University Cognitive Studies, Lund: LUCS, Brighton, 2008.

[22] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[23] R. S. Sutton, "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Machine Learning Proceedings 1990*. Elsevier, 1990, pp. 216–224.

[24] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[25] V. G. Santucci, G. Baldassarre, and M. Mirolli, "Grail: A goal-discovering robotic architecture for intrinsically-motivated learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 214–231, 2016.

[26] C. Florensa, D. Held, X. Geng, and P. Abbeel, "Automatic goal generation for reinforcement learning agents," in *International Conference on Machine Learning*, 2018, pp. 1514–1523.

[27] M. Ramicic and A. Bonarini, "Attention-based experience replay in deep q-learning," in *Proceedings of the 9th International Conference on Machine Learning and Computing*. ACM, 2017, pp. 476–481.

[28] M. Persiani, A. M. Franchi, and G. Gini, "A working memory model improves cognitive control in agents and robots," *Cognitive Systems Research*, vol. 51, pp. 1–13, 2018.

[29] D. Pacella, M. Ponticorvo, O. Gigliotta, and O. Miglino, "Basic emotions and adaptation. a computational and evolutionary model," *PLoS one*, vol. 12, no. 11, p. e0187463, 2017.

[30] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 70–82, 2010.
[31] M. Schembri, M. Mirolli, and G. Baldassarre, "Evolution and learning in an intrinsically motivated reinforcement learning robot," in *European Conference on Artificial Life*. Springer, 2007, pp. 294–303.
[32] A. Karpathy, "Reinforcejs framework," https://github.com/karpathy/reinforcejs, 2013, accessed: 2018-09-06.
[33] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

Mirza Ramicic Born in Banja Luka, Bosnia and Herzegovina on August 25th 1983 where he also obtained Bachelor and Master degree in Information Technology Engineering. He is currently a PhD candidate at Politecnico di Milano in the area of in Information Technology under the supervision of Professor Andrea Bonarini. He is a member of AI and Robotics Lab at Politecnico di Milano.

Andrea Bonarini (Milano, 1957). Laurea (Master) in Electronics Engineering (Computer Engineering area), 1984. PhD in Computer Engineering in 1989 from Politecnico di Milano. Master in Neuro-Linguistic Programming in 1993, from IIPNL.
Full professor and Chair of the PhD Program in Information Technology at Politecnico di Milano, Department of Electronics, Information, and Bioengineering.
Since 1990 he is coordinating the AI and Robotics Lab at Politecnico di Milano (AIRLab http://www.deib.polimi.it/eng/deib-labs/details/21).
He has been nominated Fellow of the Alta Scuola Politecnica (http://www.asp-poli.it) in 2012. He is among the founders of the Italian Association for Artificial Intelligence (AI*IA) and the Italian Regional Interest Group of the IEEE Neural Network Council, now Italian Chapter of the IEEE Computational Intelligence Society (Chair from 2008 to 2010). He has been from 2003 to 2006 coordinator of the Working Group on Robotics of the AI*IA. He participated since 1997 to the Robocup initiative (member of the Executive Committee from 2002 to 2010 (www.robocup.org)).
He is currently in charge of "Informatics", "Artificial Intelligence", "Robotics and Design", and "Soft Computing" courses at the Politecnico di Milano. He has given and gives courses about "Uncertainty", "Fuzzy Logic", "Soft Computing" and "Designing Interaction" within the PhD program of Politecnico. He has tutored more than 150 Laurea (Master) Theses, some ERASMUS Theses, Alta Scuola Politecnica theses, and 12 PhD Theses in the AI, Machine Learning, and Robotics fields.
He has participated and leaded several EU, national, and industrial projects. He was also representative for Politecnico di Milano in the EU Networks of Excellence on Fuzzy Systems, and on Qualitative Reasoning, and he is currently National Representative in the COST Action LUDI – Play for Children with disabilities.
Since 1989, he has realized with his collaborators and students more than 50 autonomous robots. Among the recent recognitions, he won the Kazuo Tamie Award for a robot to support therapy with autistic children.
His research interests are focusing on Human-Robot Interaction, but still include Intelligent Data Interpretation, Autonomous Robotic Agents (in particular Edutainment, Entertainment and Robogames), Affective Computing, Reinforcement Learning, and Fuzzy Systems. He has published more than 150 peer-reviewed papers on international journals, books, and proceedings of international congresses.