Proceedings of the 2007 IEEE Symposium on Artificial Life (CI-ALife 2007)

The Benefits of Boredom: an Exploration in Developmental Robotics

Scott Bolland
Key Centre for Human Factors and Applied Cognitive Psychology
University of Queensland
Queensland, Australia 4067

Shervin Emami
School of Information Technology and Electrical Engineering
University of Queensland
Queensland, Australia 4067

Abstract– Self-directed learning is an essential component of artificial and biological intelligent systems that are required to interact with and adapt to complex real world environments. Inspired by psychological and neuroscientific data, many algorithms and architectures have been proposed in the field of developmental robotics that use novelty as a training signal. Such approaches are aimed at motivating the exploration of sensory-motor contingencies for which mental models have not yet been accurately formed, driving the agent to develop task-independent competencies (such as understanding object affordances) without the need for explicit teaching. However, novelty-driven exploration on its own leads to a number of well-known problems that impede competence acquisition such as the attraction of agents to chaotic or unlearnable tasks and the temporary oversampling of aspects of the environment until they are no longer novel. This paper contributes to the field, taking insight from neuroscientific data on selective attention (particularly the temporary "boredom" associated with recently seen stimuli and a counter preference for the familiar), to propose mechanisms that may help address the noted problems relating to developmental learning in robots. Experiments conducted on an AIBO ERS-7 robotic dog demonstrate the potential of the approach.

I. INTRODUCTION

Flexible real-world problem solving often requires sensitivity to subtle task and object related features. For example, consider the task of changing a light bulb and the reasoning processes involved in selecting an object on which to stand in order to gain the appropriate height. Although the resulting selection can be expressed symbolically (e.g., "the black chair"), the reasoning process itself is sensitive to low-level task specific features such as the stability, shape, and weight-bearing characteristics of the available options. In developing artificially intelligent "thinking systems", it is doubtful that such subsymbolic sensitivities can be hand-coded, or learned through explicit tuition. Instead, learning appropriate grounded representations through interacting with the world is an important (and perhaps necessary) characteristic of artificially intelligent embodied systems.

In learning to interact with the world, it has been well argued that intrinsic reward systems that promote the exploration of the environment and sensory-motor contingencies are necessary to explain the development of a range of competencies (such as grasping, walking and interacting with objects) that occur in the absence of external reward signals (see [1] for a review). As such competencies emerge from the interplay between an agent and the environment however, the specification of potential intrinsic rewards alone does not necessarily lead to a deep understanding of the learning process. Addressing this limitation, the field of developmental robotics has recently emerged that provides a grounded platform for examining the viability of various intrinsic rewards on autonomous skill acquisition. This field not only provides insights into how self-directed learning may occur in biological systems, but provides algorithms and architectures that may be useful in training autonomous agents. To date however, the field of developmental robotics still remains in its infancy, with the vary many proposed algorithms for self-directed learning not yet being demonstrated to scale well to unconstrained real-world environments.

This paper specifically explores issues surrounding the use of novelty as an intrinsic reward signal for the self-acquisition of skills in developmental robots. As will be mentioned, novelty is a psychologically and neurologically plausible reward signal that has been widely used in the robotics community. One of the limitations of this approach however, is that rewarding only novel sensory-motor contingencies frequently leads to a number of emergent behaviors that are detrimental to learning. Through a series of simulations, this paper aims to highlight and explain some of these main detrimental emergent properties, and to propose neuroscientifically motivated extensions to generic learning architectures that may circumvent such problems. In particular, it is argued that short-term habituation to recently experienced novelty (i.e. "boredom") as well as an opposing attraction to the familiar are additional features found in biological systems that may address some of the main limitations associated with using novelty as an intrinsic motivator.

II. BACKGROUND

The factors that motivate an organism to interact with its environment and select certain actions over others has been a subject of much debate over the past century. Early drive theorists argued that behavior is motivated by the desire to reduce tension caused by unmet biological needs such as hunger, thirst or sex (e.g., [2]). However, since that time, psychologists have recognized that drive reduction theories

are insufficient to account for a wide range of human behaviors such as visual exploration, grasping, walking, language, the exploration of novel objects, and the general ability to exploit and manipulate the environment [1]. Instead, it has been argued that there exist additional intrinsic rewards that make pleasurable the vast amount of learning that is required to effectively interact with the world. Such additional drives that have been proposed include the need to explore, the need to effect a stimulus change in the environment, and the need to interact effectively with the environment [1,3,4].

Recently, the field of developmental robotics has emerged that, in conjunction with neuroscientific data, has given much insight into how competencies (i.e. sets of effective interactions with the world) can emerge in the absence of external rewards. One such factor that has been found to be an effective reward signal is stimulus novelty (e.g., [5-11]). That is, the discrepancy between sensory stimuli and that predicted by an internal world model can be used as a training signal to promote the exploration of aspects of an agent's environment for which a mental model has not yet accurately formed. Although generic in nature, variants of this approach have been successfully applied to the acquisition of a range of competencies, such as the learning of object permanence, object affordances and language (e.g., [10,,11]). Interestingly, the use of novelty as a reward signal automatically promotes exploration, manipulation, and mastery; the main intrinsic motivators identified by psychologists (as discussed earlier).

Consistent with the algorithms prevalent in developmental robotics, there is much neuroscientific evidence to suggest that novelty acts as an important intrinsic reward signal in primates. In particular, dopamine neurons of the ventral tegmental area and substantia nigra have been shown to elicit the same response to novel stimuli as they do to primary rewards such as food [12]. It has long been established that such cells are central to the processing of rewards that shape overt behavior (see [12] for a discussion). FMRI studies suggest that stimulus novelty is detected in the hippocampus, where temporal predictions are matched against current sensory stimuli (e.g., [13,14]). Apart from exciting cells that help shape long-term behavior [15], the hippocampal novelty signal is also believed to directly trigger the orienting response that directs attention towards the novel stimuli [16].

Although the utility of novelty-driven exploration and learning has been demonstrated in the area of developmental robotics on a range of tasks, there exist several unresolved issues that prevent its scalability to complex real-world environments. Firstly, it is commonly highlighted that using novelty as an intrinsic reward signal will fail in environments where there are regions of unlearnable contingencies as the agents will be drawn to these areas (e.g., [11,17]). For example, watching passing traffic would be innately interesting to such agents, as it is impossible to predict what type or color of car would appear next. To circumvent such problems, and consistent with theories of human behavior, it is assumed that exploration should occur somewhere between situations of complete familiarity (boredom) and complete unfamiliarity (e.g. chaos) (e.g., [3,5]). Although some algorithms have been proposed to promote exploration of tasks in which learning is actually occurring (e.g., [11,17]), as yet, they have not been demonstrated to function in large-scale complex environments. Furthermore, such approaches are generally not neuroscientifically inspired, and thus provide little insight or explanation as to how self-directed task learning occurs in humans.

In addition to the obvious side-effects of pure novelty-driven exploration, there exists other related but non-obvious emergent behaviors that equally impede the scalability of the approach. One such example, and the main focus of this paper, is the phenomenon of oversampling; the fixating of an agent on a specific task (at the exclusion of others) until it has been successfully mastered (e.g., [5,11]). Such an oversampling of a sensory-motor contingencies may lead to the consequence of the action becoming prematurely "familiar", and the action no longer being displayed. As will be demonstrated in this paper, although not necessarily a problem for simple environments, oversampling can be impede the learning of more complicated tasks. For example, many self-acquired competencies require a form of skill scaffolding, in which a simpler task (e.g., moving your hand in the direction of your gaze) needs to be performed in a wide range of contexts before its utility is discovered (e.g., the ability to touch a visible object). This paper investigates (through a series of simulations), the nature and cause of oversampling, and provides neuroscientifically motivated mechanisms that may provide insights into how the problem may be addressed. Implications as to how the proposed solution may be extended to help agents be directed away from unlearnable tasks is also discussed.

III. SIMULATIONS

A. The Task Environment

The aim of the project described in this paper is to explore the phenomenon of oversampling in self-directed learning, and examine the efficacy of using short-term boredom (described and justified later) as a potential biologically motivated solution. The environment chosen for this purpose is an extension of that described by [11] that was used to explore affordance learning using Intelligent Adaptive Curiosity (IAC) (an algorithm that uses detected decreases in novelty to promote the exploration of learnable tasks).

The original environment described in [11] was fairly simplistic, consisting of an AIBO ERS-7 that learned to interact appropriately with an elephant ear toy (that could be bitten), a suspended toy (that would oscillate when "bashed") and a second "adult" AIBO (that imitated the sounds made by the first robot). The main limitation of this environment however, was that the AIBO was stationary, only being able to move its head, front legs and jaw in a set of prespecified motions. Given the types of representations that were used and the fact that the objects of interest were also stationary, the affordances that were acquired were location specific. That is, learned affordances included such facts as that looking and "bashing" to the leftmost position would result in

oscillation (as this was the location of the hanging object), rather than more general location invariant rules of interaction. Although oversampling looked to have occurred in this environment (i.e. certain actions became prevalent at different times during learning), it was not deemed to be a problem, as it was interpreted as an indication that the various affordances were being explored in order of their complexity (consistent with the IAC algorithm).

For the simulations described in this paper, we extended upon the above environment to include a more natural task in which affordances would need to be learned in a location invariant fashion (as opposed to mapping each affordance to a specific motor command). This was achieved by allowing AIBO to turn both its head and body, so that objects could appear in any location. Thus, to acquire the general skill of hitting hanging objects, AIBO would need to learn what motor action corresponded to hitting in each potential direction of current gaze (provided that a hanging object was visible). From previous experience, we were aware that such a task would be difficult to learn for algorithms that were susceptible to the oversampling problem. Specifically, as oversampling promotes the exploration of only specific actions at a time at the exclusion of others, it was predicted that all the actions corresponding to a general skill (such as hitting a hanging object in any direction) would be unlikely to be exhibited together. Thus it was postulated that such an environment would allow us to examine the problem of oversampling in self-directed learning and explore potential solutions.

The extended environment that was used for our simulations consisted of an AIBO ERS-7 placed on a swivel chair that it could turn either left or right, with all relevant objects being placed at arms length (shown in Fig. 1). The objects that the robot could interact with included a hanging toy that oscillated when "bashed", a piece of foam at eye level that could be bitten, and a "Dora the Explorer" musical toy that would play a tune when one of its buttons was pressed. Actions performed by the robot were sequential, chosen from fourteen preset motor commands: the AIBO could face a random direction (by swiveling left or right, or turning its head in one of 6 directions); it could perform a bashing movement with its front legs in one of six directions; it could perform a button pressing motion in one of six directions; or it could bite in the direction it was facing. These specific actions allowed the robot to randomly explore its environment, and perform movements that would allow it to interact with the various objects.

The sensory inputs to the system were chosen to allow the AIBO to acquire enough information about the environment and its internal state to select an appropriate action, and to detect what effect (if any) the action had on the environment. The inputs included AIBO's current head position (of which there were 6 preset angles), whether or not a green, pink or yellow object was visible (corresponding to the hanging toy, the bitable foam and the musical toy respectively), whether or not its hand was seen moving, whether or not there was additional motion in the environment, whether or not an object was bitten on the last iteration, and whether or not a tune was heard. As will be described later, the ability for the AIBO to see its own hand moving was included to promote actions in the direction of its gaze, facilitating affordance learning (as it is only when the AIBO was facing an object that the effects of its actions on the object would be detected). The second motion detector was used to detect the presence of an oscillating object in its field of view (i.e. caused by the hanging object), but would also be activated whenever AIBO moved its head (as this would result in a major change in the visual field).

Fig. 1. The Task Environment. An AIBO ERS-7 is mounted on a swivel chair surrounded by a hanging toy, a bitable piece of foam and a "Dora the Explorer" musical toy that it can learn to interact with.

B. The Agent Controller Model

The environment chosen for the simulations was selected so that novelty could be used as an appropriate reward signal, allowing the use of relatively simple architectures and algorithms for learning. Given the specific environment, the only action that would lead to an unpredictable consequence was that of turning, as the representations used did not allow predictions as to what objects, in any, would become visible as a result. As random exploration is a desirable default behavior for the simulation, such residual "novelty" is beneficial rather than detrimental. Thus rather than using a fairly complex algorithm such as IAC that promotes the exploration of learnable regions of the problem space, a more simple approach using novelty as a reward can be used so that the root cause and potential solutions to oversampling can be studied in isolation.

Consistent with previous "curious" model-building control systems (e.g., [8,11]), the agent used in our simulations contained two separate learning modules: an action selection module (that selects an appropriate action given the current sensory input) and a prediction machine (that generates an expectation of the resulting sensory input). Similar to [8] both the action selection module and prediction module were implemented within a common three-layered neural network, taking a sensory input vector and action vector as input (see Fig. 2). The aim of the prediction module was to predict the resulting sensory state given the current state and action being performed. For example, given that there was a hanging object in view, the network would learn that hitting in the

direction of the current head position would cause the object to oscillate. After each action was performed, the actual resulting sensory state was used to train the network to refine its predictions.

In contrast to the prediction module, the action selection module output a single value that reflected the utility of performing a specific action given the current sensory state. In selecting the next action to perform, all 14 of the possible actions were fed in series through the module, with their corresponding utilities being calculated. The next action to perform was selected using Softmax action selection with a Boltzman Distribution [18]. This procedure adds Gaussian noise to each of the selection strength values, choosing the action with the highest resulting value. Using this approach, actions occur with a probability that is a graded function of their estimated utility. After each action was performed, the action selection module was trained to better predict the error associated with the given action in the prediction module (the highest absolute difference between the predicted and actual sensory states on any one feature). Using this discrepancy as a novelty signal, the agent gradually became biased to explore context-specific actions for which a mental model had not yet accurately formed, leading to the active exploration of the task environment.

Specifically, the input layer of the implemented network consisted of a sensory input array containing 9 units and an action pathway containing 14 input nodes (see Fig. 2). The sensory input array consisted of 6 nodes representing the head position (using a local coding scheme), with the remaining 3 nodes being the binary feature detectors for each type of object. The action array contained a single node for each of the 14 actions that the AIBO was able to perform. When the "random exploration" action was chosen, one out of 8 possible motions would result with equal probability, corresponding to a swivel in the chair either left or right, or a head turn to one of 6 angles. The output of the prediction pathway was a vector containing the 7 binary features representing external properties that may be affected by the action, including the presence of objects at the new location, and the generation of noise or motion. To simplify the learning process, and to better understand the representations learned by the network, the input to hidden layer weights were fixed to bind various

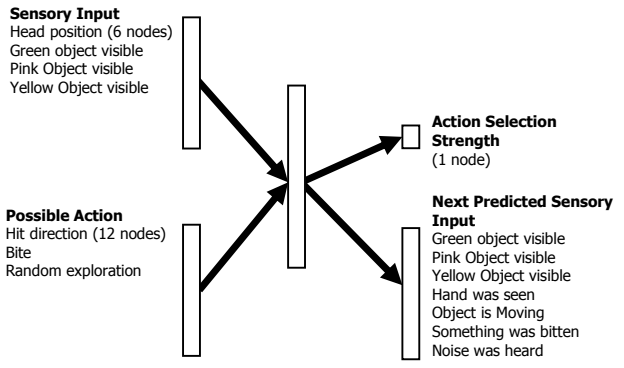

Fig. 2. The Neural Network Controller. Each separate input or action is represented by a separate network node (except where specified above).

variables together in a way that would allow the task to be learned within a single layer of trainable weights (i.e. the hidden to output weights). Firstly, there were a set of hidden nodes corresponding to the outer product of the complete network input vector with itself. This set of nodes thus included a binding of head and hitting position, of which there were 36 unique combinations (i.e. each combination being represented by a unique node). A portion of these nodes (representing combinations in which the hitting action is in the same direction as the head) could be used to directly predict that movement will occur in the visual field (i.e. the hand would be seen), thus providing a useful representation to learn this mapping. To facilitate the learning of object affordances, a similar representation was used, consisting of the outer product of the input vector with itself and the object identification vector (containing binary object detectors for the three possible objects). Thus, there was unique node representing when a particular hitting action paired with a head position occurred in the presence of a specific object. Such representations contain nodes that could be directly used to predict if oscillation, noise or successful biting would occur. Of course however, such representations also yield many features that are not useful for predicting the occurrence of object interactions (such as nodes corresponding to when actions and head movements are uncoordinated), thus still requiring a high degree of learning to master the task.

As the input to hidden weights were fixed, the trainable region of the network consisted of a layer of weights joining the hidden unit representations with the output layer. This pathway was implemented as a single layered neural network using standard sigmoid activation functions, and training algorithms. It should be noted however that the biases on the outputs of the "prediction module" were initially preset to a value of -3 (plus Gaussian noise), resulting in a predicted output of all sensory features close to 0. This reflected the fact that the detection of the chosen features (such as movement or specific objects) are rare and should be considered initially "novel."

C. Simulation 1: Learning Hand-Eye Coordination

The aim of the first experiment was to explore the efficacy of the simple agent at learning the given tasks. As mentioned earlier, due to the oversampling problem, it was predicted that general skills (such as hitting objects irrespective of angle) would be unlikely to emerge. Instead, it was postulated that different actions associated with a skill would be explored and exhibited at different times.

The general environment chosen for the set of simulations is interesting in that main affordances that can be learned (i.e. hitting, biting or pressing corresponding objects), are somewhat needle-in-a-haystack type problems. That is, for example, a hanging object lies in roughly 1/16 of the circular area that the AIBO can face, with there being a 1/12 chance that a random hitting action will be appropriate (i.e. coordinated with the height and direction of the object). Thus, in this specific case, a successful interaction with the object will only occur in 1/192 of random trials (which of course

would be far less probable if one were to scale up the environment). In order to facilitate the learning of such tasks, other subtasks were added that acted as a form of scaffolding, rewarding behaviors that were useful in acquiring the more difficult tasks.

The main subtask that was added to facilitate learning was that of hand-eye coordination, biasing the emergence of hitting or pressing motions in the direction of current gaze (irrespective of the presence or absence of an object). With such a bias present, the probability of successful interactions with a hanging object are greatly increased (as the robot would be likely to attempt either pressing or hitting in the correct direction), facilitating affordance learning. Such scaffolding can be readily seen in human infants, where many initial behaviors do not have a direct effect on the world, but are useful in learning more complicated tasks. For example, newborn infants have been shown to spend up to 20% of the time touching their hands to their face [19], with such exploratory motions being viewed as a way to learn to control the dynamics of their bodies [20]. The bias for learning hand-eye coordinated movements was implemented by using a feature that would detect the movement of the hand as a result of the previous action. As hitting in the direction the robot was facing would lead to the surprising appearance of the hand, this action would be reinforced.

In the first experiment described below, the aim was to explore the efficacy of the agent at learning the simpler task of hand-eye coordination, before scaling up to the more difficult environment. For this experiment, objects were placed just out of reach (so that affordances would not be learned), with the only actions permitted being hitting (i.e. in one of 6 directions), and random turning. For this and the following experiments, a relatively slow learning rate of 0.01 was used to train the prediction network (to prevent catastrophic interference from recent events), and a faster rate of 0.1 to train the action selection network (as this is simply attempting to mirror the error from the first network). A temperature value of 0.03 was used for the Softmax selection, resulting in a strong bias to exhibit actions with a higher associated error.

Over 100 runs of 15000 iterations, using the simple architecture described earlier, the simulation consistently demonstrated oversampling, with specific behaviors being acquired and exhibited in a serial rather than parallel manner. As shown in Fig. 3 (in terms of the selection strength of various actions for an example trial), although initial actions were random, random head turning was quickly selected for as it would result in the "novel" feature of "movement being seen". This action was performed at the exclusion of other actions from iterations 300-1000. Once the consequence of this action was well predicted by the network however, the selection strength dropped, allowing the exploration of other actions. Rather than hand-eye coordination being learned in parallel for all directions, the task was learned in series, focusing on a particular angle at a time. For example, between iterations, 1500-1800, AIBO repetitively hit in direction 5 (the direction that it was facing), preferring this action over all others. Once the consequence of this action

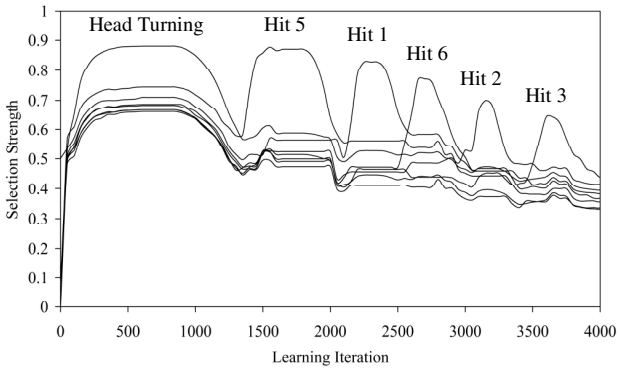

Fig 3. Selection strength of various actions demonstrating a serial exploration of competencies. After an initial phase of head turning, coordinated hitting became temporarily strengthened in a number of different directions.

was well predicted, the system then focused on coordinating a new angle. As shown in figure 4, the above pattern of acquisition was typical, with the agent only ever performing coordinated hitting across 40% of angles at a time (collapsed across 100 trials).

The selective mastering or exploration of one task at a time found in our simulations is not uncommon in the literature, being reported in many developmental robotics simulations (e.g., [5, 11]). Rather than being a noted problem however, such oversampling has been interpreted as a form of staged learning that could potentially account for human skill acquisition [11]. However, in such simulations, each affordance to learn mapped directly onto a single action (for example, in [11], hitting and looking in the leftmost direction would move the hanging object). In contrast, in our simulations, learning a general skill or affordance requires the mapping of a number of distinct motor commands. For example, to successfully exhibit general hand-eye coordination, one must learn the motor commands associated with looking in each direction. As highlighted by our first simulation, if angles are learned and mastered in a serial manner, the overall skill will never emerge. Thus, the phenomena of oversampling does not seem to be a true

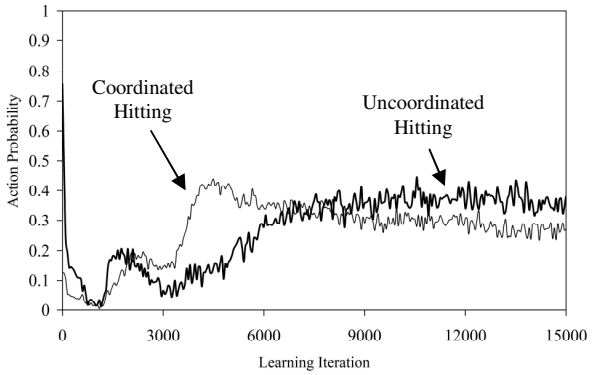

Fig 4. Overt behavior of the simple agent averaged across 100 trials. Coordinated hitting only occurs synchronously in a couple of directions between iterations 3000-6000, with the system then regressing to random behavior (uncoordinated hitting).

reflection of human learning, but instead may actually be detrimental to skill acquisition.

D. Simulation 2: The Implementation of Boredom

As mentioned in [5], the nature of the oversampling problem is quite straightforward to understand. Initially, all salient events (such as seeing a hand or causing an object to oscillate), will be unpredicted. Each time such an event occurs, the corresponding action will be reinforced to occur more frequently in the given context (i.e. the output of the action selection network will be trained to give a higher value, to more closely match the prediction error). This increase in selection strength will itself increase the rate of sampling of the behavior, bootstrapping the process until the action is performed at a high frequency (potentially at the exclusion of many other important behaviors). Once the consequence of this action is well-predicted however, the intrinsic reward once again diminishes, with the agent getting "bored" of exploring this sensory-motor contingency, moving on to other tasks.

Apart from tasks such as ours in which oversampling prevents the exhibition of general competencies, oversampling in a developmental agent may be problematic for several other reasons. As stated previously, oversampling is the temporary fixation on a task at the exclusion of others until it is completely mastered. However, in the real-world there are many tasks which cannot be completely mastered (such as predicting the color of the next car that will appear out of a tunnel). For such chaotic behaviors, the error in prediction will always be high, which may result in an agent getting permanently stuck in the exploration of the task. A separate problem is that, in neural networks at least, if learning episodes are not interleaved, catastrophic forgetting of older information can occur [21]. Thus in such systems, it is likely that previously learned knowledge about sensory-motor contingencies may be corrupted if it is acquired in series and not consistently reexamined over time.

In order to address the above limitations, what is required is an additional feature of the learning agent that explicitly prevents oversampling and promotes the exploration of many activities in parallel. Our suggested approach to achieving the parallel acquisition of tasks is motivated by studies of human attention and perception that demonstrate that attention itself is directed toward novel (as opposed to recently experienced) stimuli (e.g., [16,22]). For example, in infant "habituation" experiments, when repetitively shown pairs of pictures, infants will spend more time looking at new as opposed to recently experienced images [23]. There is much evidence to suggest that this novelty signal is generated in the hippocampus, where fast context-dependent learning can occur [24]. Relevant to our study however, is that the fact that stimuli in a given context become rapidly familiar (over a few presentations), no-longer eliciting the novelty signal [25]. This habituation is viewed to be short-lived and context dependent [16]. It is our belief that such short-term boredom that helps direct attention away from the current stimuli, may be a viable and psychologically plausible mechanism for overcoming the phenomena of oversampling.

The aim of this paper is not to argue the cause or loci of temporary attentional habituation, but rather, to explore the utility of such habituation in the prevention of oversampling, Simulation 2 examines the architectural addition of attentional habituation into the simple controller network described earlier. In implementing such habituation, a number of assumptions are made: firstly, that habituation of the novelty response is fairly short lived, and secondly, that habituation occurs to recently experienced patterns of sensory activity rather than individual features (for example, a yellow banana would still be viewed as novel and interesting, following inhibition to a yellow ball).

In the following model, attentional habituation was implemented using a simple mechanism. Firstly, the sensory array (a binary vector) for the last 5 patterns was held in memory (i.e. a short-term storage of information). The probability of looking away (i.e. for attention to be directed elsewhere), was calculated as a linear function of short-term familiarity of the current sensory input. Specifically, if the current sensory pattern occurred once in short-term memory, there was a 30% chance of looking away, if it occurred twice, there was a 60% chance, etc. (reaching a threshold of 100% for four or more occurrences). Redirecting attention away from the current stimuli was achieved through the triggering of the "random exploration" action leading to either random head movement or the turning of the robot on the swivel chair.

Apart from implementing a form of short-term boredom to promote exploration, a second modification was made to prevent the system regressing to random behavior. That is, in the previous simulation, the intermittent reinforcement of seeing unpredicted objects during random exploration was not sufficient for this action to be significantly strengthened. As a result, after coordinated hitting was explored, the system would once again regress to uncoordinated hitting (see Fig. 4). To circumvent this problem, and allow intermittent rewards to better shape behavior, different learning rates on action selection strengths were used for when the actual sensory error was higher or lower than what was predicted by the action selection network. Specifically, when the actual error was higher than predicted (corresponding to a "surprising event") the action would be highly reinforced (by using a learning rate of 0.1), whereas if the error was lower than expected, the action would more slowly decay (using a learning rate of 0.01). This differential learning strategy used to promote intermittently reinforced behaviors was also biologically motivated, reflecting the fact that dopamine neurons send out a strong positive signal at the presence of an unpredicted reward, and a weak negative signal when an anticipated reward does not occur [12].

Apart from the adaptations described above, using exactly the same parameters and method for the first experiment, the simulation was retested. As can be seen in Fig. 5 (showing overt behavior collapsed across 100 simulations), oversampling no longer occurred, with coordinated hitting being learned across all angles. However, over time, as would

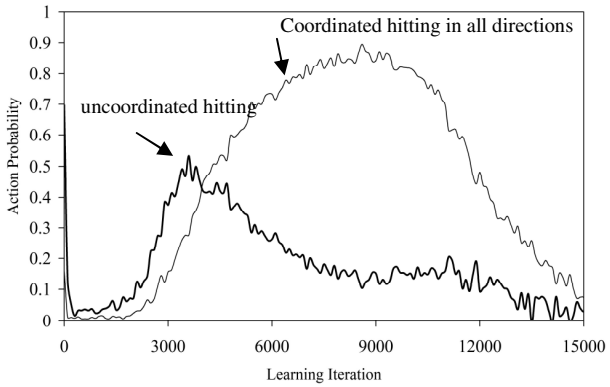

Fig 5. Parallel exploration of competencies across 100 trials. After an initial phase of head turning, coordinated hitting was exhibited in all directions.

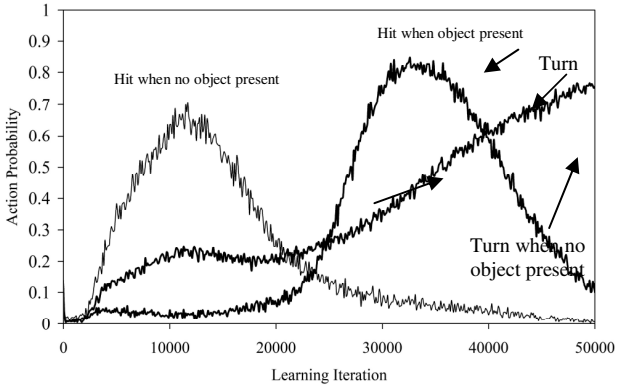

Fig 6. Scaffolded task learning. After an initial phase of coordinated hitting, this behavior regresses to random exploration around iteration 12000, but continues hitting when objects are present.

be predicted, the consequences of coordinated hitting became well-learned, with the error (and related selection strength) decreasing over time. As a result, this behavior diminished. Unlike the previous experiment however, random exploration was reinforced, becoming the default action over time.

E. Simulation 3: Scaffolded Learning

The aim of simulation 3 was to explore the self-acquisition of affordances, using novelty as a reward signal. For this simulation, all of the objects specified earlier (i.e. a hanging toy, a bitable piece of foam, and a musical interactive toy) were added to the environment. As the objects were only located in one location in AIBO's circular environment, random movement (i.e. head or body rotation) would result in a small probability that each object would be visible. As mentioned earlier, to facilitate learning of this needle-in-a-haystack type problem, hand-eye coordination was reinforced (through the initially unpredicted appearance of the hand), so that the degrees of exploration would be reduced when an object was present. That is, when an object was present, the robot would be biased to either attempt a pressing or bashing object in the direction of its gaze as opposed to the other 5 possible angles. Thus, it was postulated that there would be an increased probability that the behavior would be appropriate for the object. In contrast to pressing and bashing motions, biting did not require scaffolding, as the action was automatically coordinated with current gaze, and as such was an easier task to master.

As show in Fig. 6 (depicting the average behavior across 100 trials), using the same parameter settings as the previous experiment, scaffolded learning occurred naturally in the system without any additional modifications. As seen in this figure, the more basic task of coordinated hitting was learned quickly and early, but started to extinguish around 12000 iterations. In contrast, when an object was in view, appropriate hitting or pressing actions (which were initially facilitated by the coordinated hitting behavior), continued to be promoted even up to iteration 50000. During this phase, the biting affordance was also displayed heavily, although peaking earlier, due to the relative simplicity of the sensory-motor contingency being learned.

In summary, this simulation exhibited distinct phases of behavior, firstly, learning to coordinate hand and head motions (irrespective of whether or not an object was present), and then later exhibiting only object appropriate actions (i.e. exploring the environment until an object was detected, and then interacting with it directly until sensory inhibition forced the agent to move on). Such a phase of object appropriate interactions was temporary however, as the consequences of each action could be accurately learned over time, leading to a final stage of exploratory behavior in which all objects were disregarded.

IV: Discussion

In self-directed learning, an agent actively explores the environment in order to build an accurate world model and to develop task-independent competencies. Much neuroscientific and psychological evidence has been presented in the literature to suggest that novelty may act as an intrinsic motivator that facilitates the process of skill acquisition by motivating an agent to explore aspects of the environment for which an accurate mental model has not yet formed. As mentioned in this paper however, novelty by itself is not sufficient to explain the acquisition of competence, as it may lead to several emergent behaviors that impede learning. For example, exploration driven by novelty can lead to a form of bootstrapping in which a specific action is oversampled at the exclusion of others. Likewise, novelty seeking may cause the agent to prefer chaotic and unpredictable regions of its task environment.

What is evident in current research (both in the practical application of developmental robotics and in understanding biological systems) is that mechanisms that promote the exploration of only moderate levels of novelty are required. This paper attempts to propose a solution to the various problems associated with novelty-driven search through gaining inspiration from psychological and neuroscientific research. In particular, we argue that much insight may be

gained from understanding the role of the hippocampus on attention and selection, and how dopamine neurons encode rewards and process intermittent reinforcement. In particular, it is evident that there are short-term inhibitory effects on attention to recently experienced stimuli that may help prevent an agent from oversampling the interactions with a particular object or environment. In this paper we demonstrate that a relatively simple account and implementation of this process can lead to the scaffolded learning of competencies that are otherwise unlearnable in agents of the similar complexity.

With respect to possible extensions to this work we note that there exist other aspects to the way in which the brain guides attention that may also be highly relevant in understanding developmental learning. In particular, as argued by [20], there may be two quite distinct pathways that guide attention: a pathway via the hippocampus that directs attention towards novel stimuli (habituating rapidly as emulated in our simulations), and a second pathway through the cortex that draws attention towards familiar stimuli. As a result of these parallel pathways, the brain may be biased to attend to novel aspects of familiar situations and environments. Thus, it may be the case that understanding the interactions between these various pathways may hold the key to understanding how to bias an agent to explore only moderate levels of novelty; an essential behavior in the general acquisition of competence. This proposal is in contrast with approaches such as [11] and [17] in which regions of the environment in which learning is predicted to be likely are explicitly calculated and used as a reinforcement signal; a mechanism for which no direct neural mechanism has been identified.

In conclusion we argue that possible solutions to current salient problems in the area of developmental robotics might be found through a greater understanding of the role of the cortex and hippocampus on directing attention, and the role that the dopaminergic system plays in the processing of rewards and novel stimuli.

ACKNOWLWDGMENT

This research was supported by ARC Linkage grant No. 433-5248-03 and NSRSF grant No. 1-22-5248-75. The authors wish to thank the various members of the ARC Centre for Complex Systems for their feedback and support.

REFERENCES

[1] R.W. White. "Motivation Reconsidered: the Concept of Competence," Psychol Rev., vol. 66, pp. 297-333, Sep. 1959.

[2] C.L. Hull, Principles of Behavior. New York: Appleton-Century-Crofts, 1943.

[3] D. Berlyne, Conflict, Arousal, and Curiosity. New York: McGraw-Hill, 1960.

[4] E.L. Deci, and R.M. Ryan, "The What and Why of Goal Pursuits: Human needs and the Self-Determination of Behavior," Psychol. Inq., Vol. 11, No. 4, pp. 227-268, 2000.

[5] A.G. Barto, S. Singh and N. Chentanez, "Intrinsically Motivated Learning of Hierarchical Collections," Proc. of Inter. Conf. on Developmental Learning, 2004.

[6] D. Blank, D. Kumar, L. Meeden and J.B. Marshall, "Bringing Up Robot: Fundamental Mechanisms for Creating a Self-Motivated, Self-Organizing Architecture," Cybernet. Syst., vol. 36, n 2, pp. 125-150, March 2005..

[7] X. Huang and J. Weng, "Novelty and Reinforcement Learning in the Value System of Developmental Robots," Proceedings of the Second International Workshop on Epigenetic Robotics. Modeling Cognitive Development in Robotics Systems, pp. 47-55, 2002.

[8] J. Marshall, D. Blank and L. Meeden, "An emergent framework for self-motivation in developmental robotics," Proceedings of the Third International Conference on Development and Learning, pp. 104-111, 2004.

[9] J. Weng and Y. Zhang, "Developmental robots - a new paradigm," Proceedings of the Second International Workshop on Epigenetic Robotics. Modeling Cognitive Development in Robotics Systems, pp. 163-74, 2002.

[10] Y. Chen and J. Weng. "A Case Study of Developmental Robotics in Understanding Object Permanence" BICS2004 Aug29-Sept 1, 2004

[11] P.Y. Oudeyer and F. Kaplan, "The discovery of communication," Connect. Sci., vol. 18, n. 2, pp. 189-206, 2006.

[12] W. Schultz, "Predictive reward signal of dopamine neurons," J Neurophysiol, Vol. 80, pp.1-27, 1998.

[13] D. Kumaran and E.A. Maguire, "An Unexpected Sequence of Events: Mismatch Detection in the Human Hippocampus," PLoS Biol., vol4, n. 12: e424, 2006.

[14] S. Yamaguchi, L. A. Hale, M. D'Esposito, and R. T. Knight, "Rapid Prefrontal-Hippocampal Habituation to Novel Events," J. Neurosci., vol. 24, n. 23, pp. 5356 – 5363, 2004.

[15] J.E. Lisman and N.A. Otmakhova, "Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine," Hippocampus, vol. 11, pp. 551–568, 2001.

[16] O.S. Vinogradova, "The hippocampus and the orienting reflex," In E.N. Sokolov and O.S. Vinogradova (eds.), Neuronal mechanisms of the orienting reflex. Hillsdale, NJ: Erlbaum, pp. 128-154, 1975.

[17] J. H. Schmidhuber, "Curious model-building control systems," In Proc. Intl. Joint Conf. on Neural Networks, pp. 1458-1463, 1991.

[18] R.S. Sutton and A.G. Barto, Reinforcement Learning. An Introduction. Cambridge, MA: MIT Press, 1998.

[19] A.F. Korner and H.C. Kraemer, "Individual Difference in Spontaneous Oral Behavior in Neonates." In J. Bosma (ed.) Proceedings of the 3rd Symposium on Oral Sensation and Perception, pp. 335-346, 1972.

[20] A. Smitsman and R. Schellingerhout, "Exploratory behavior in blind infants: how to improve touch?" Infant Behavior and Development, Vol. 23, pp. 485-511, 2000.

[21] M. McCloskey and N. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," In G. H. Bower (ed.) The Psychology of Learning and Motivation: Vol. 24, pp. 109-164, NY: Academic Press, 1989.

[22] E.N. Sokolov, "The neuronal mechanisms of the orienting reflex," In E.N. Sokolov and O.S. Vinogradova (eds.), Neuronal mechanisms of the orienting reflex. Hillsdale, NJ: Erlbaum, pp. 217-235, 1975.

[23] R.L. Franz, "Visual experience in infants: Decreased attention to familiar patterns relative to novel ones," Science, vol. 146, pp. 668-670, 1964.

[24] J.E. Lisman and N. Otmakova, "Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine," Hippocampus. Vol 11, pp. 551-558, 2001.

[25] S. Yamaguchi, L. Hale, M. D'Esposito, and R.T. Knight, "Rapid prefrontal-hippocampal habituation to novel events," J Neurosci, Vol. 24, n. 23, pp. 5356-5363, 2004.