Nurturing promotes the evolution of reinforcement learning in changing environments regardless of instincts

Journal Title
XX(X):1–27

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Syed Naveed Hussain Shah[1,2], Ingo Schlupp[3], and Dean F. Hougen[1]

Abstract

To understand why learning is central to the behavioral repertoires of some organisms yet is largely or completely absent in others, we need to understand the evolution of learning and instincts. The evolution of learning is also an important topic in robotics and machine learning, as evolutionary approaches to developing learning algorithms have shown great promise. The evolution of reinforcement learning is particularly relevant to both biology and artificial intelligence because reinforcement learning allows an agent to directly interact with its environment and learn complex behaviors based on evaluative feedback—how well did the agent's actions actually work out in practice? This makes reinforcement learning potentially highly beneficial. However, reinforcement learning requires exploration of unfamiliar situations, which necessarily involves unknown and potentially dangerous or costly outcomes. The interplay between potential benefits and potential costs means that reinforcement learning is likely to evolve in some niches but not in others. Here we explore whether nurturing—one individual investing in the development of another individual with which it has an ongoing relationship—helps to create a niche in which reinforcement learning thrives. The results show that nurturing promotes the evolution of reinforcement learning in various changing environments.

Keywords

reinforcement learning, nurturing, evolution, robotics

1 Introduction

Artificial evolution of learning (as summarized by Soltoggio, Stanley & Risi, 2017) seems a promising approach to greater robot intelligence, given the potential of evolutionary robotics (Bongard, 2011; Bongard, 2013; Eiben, 2014; Haasdijk, Bredeche, Nolfi & Eiben, 2014; Vargas, Di Paolo, Harvey & Husbands, 2014; Doncieux, Bredeche, Mouret & Eiben, 2015; Eiben & Smith, 2015; Nolfi, Bongard, Husbands & Floreano, 2016; Silva, Correia & Christensen, 2016; Husbands, 2017) but the question is what environments promote the evolution of learning? The key hypothesis guiding this research is that there is a virtuous cycle between nurturing and the evolution of learning, where each promotes the other, as illustrated in Figure 1 (Woehrer, Hougen, Schlupp & Eskridge, 2012).

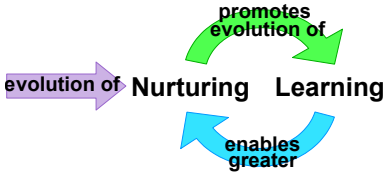


Figure 1. Hypothesized virtuous cycle between nurturing and learning with the evolution of nurturing as the entry point.

[1]School of Computer Science, University of Oklahoma
[2]Microsoft Corporation
[3]Department of Biology, University of Oklahoma

Corresponding author:
Dean F. Hougen, Robotics, Evolution, Adaptation, and Learning Laboratory (REAL Lab), School of Computer Science, University of Oklahoma, Norman, OK 73019

Email: hougen@ou.edu

Developing truly autonomous robots is an important research challenge. Currently, even highly sophisticated robots are tightly controlled by human operators (Bekey, Ambrose, Kumar, Lavery, Sanderson, Wilcox, Yuh & Zheng, 2008; Chen, Chen & Chase, 2009; Sukhatme, 2009; Dudek & Jenkin, 2010; Huntsberger, Keegan & Brizzolara, 2010; Huntsberger, Keegan & Brizzolara, 2011; Siegwart, Nourbakhsh & Scaramuzza, 2011; Kapach, Barnea, Mairon, Edan & Ben-Shahar, 2012; Marques, De Almeida, Armada, Fernández, Montes, González & Baudoin, 2012; Theodoridis & Hu, 2012; Hexmoor, 2013; Liu & Nejat, 2013; Mirbagheri, Baniasad, Farahmand, Behzadipour & Ahmadian, 2013; Doelling, Shin & Popa, 2014; Maurer, Steinbauer, Lepej & Uran, 2014; Zuzánek, Zimmermann & Hlaváč, 2014; Behere, 2015; Fonseca & Pontuschka, 2015; Murphy, 2015; Yanco, Norton, Ober, Shane, Skinner & Vice, 2015; Kronreif, 2016; Taylor & Riek, 2016).

To be autonomous, robots need traits that allow them to respond adaptively to changes in their environments. An extremely successful template for this is found in the biological world. Many animals can independently process enormous amounts of information and use it autonomously for their benefit. Learning has evolved to play a central role in the life of many animals, indicating the adaptive value of learning overall (Harley & Smith, 1983; Bolles & Beecher, 1988; Stephens, 1991; Feldman & Laland, 1996; Moore, 2004; Dukas, 2013; Snell-Rood, 2013; Dridi & Lehmann, 2014; Lotem & Biran-Yoeli, 2014). By contrast, the learning capabilities of even our most advanced robots are extremely modest (Argall, Chernova, Veloso & Browning, 2009; Peters, Morimoto, Tedrake & Roy, 2009; Kober & Peters, 2012; Kober, Bagnell & Peters, 2013; Kormushev, Calinon & Caldwell, 2013; Peters, Kober, Mülling, Krämer & Neumann, 2013; Chernova & Thomaz, 2014; Ghanbari, Vaghei, Noorani & Reza, 2014; Kober & Peters, 2014a; Pagliuca & Nolfi, 2015; Amarjyoti, 2017; Ingrand & Ghallab, 2017; Polydoros & Nalpantidis, 2017).

Learning was explored early on in robotics (Walter, 1951; Walter, 1953) and it has been argued that true autonomy requires learning (e.g., Bekey, 1996; Breazeal, 2004; Commuri, Fierro, Hougen & Muthuraman, 2004; Ott & Ramos, 2013; Sales, Santos, Sanz, Dias & García, 2014). Moreover, learning has allowed for the development of robotic systems that have outperformed non-learning systems on complex tasks from robot soccer (Stone & Veloso, 2000) to autonomous driving (Thrun et al., 2007) to object picking and stowing (Hernandez et al., 2016). However, many challenges to robot learning remain, including the development of appropriate learning mechanisms for complex tasks, environments, and representations (Argall et al., 2009; Peters et al., 2009; Kober & Peters, 2012; Kober et al., 2013; Kormushev et al., 2013; Peters et al., 2013; Chernova & Thomaz, 2014; Ghanbari et al., 2014; Kober & Peters, 2014a; Pagliuca & Nolfi, 2015; Amarjyoti, 2017; Polydoros & Nalpantidis, 2017).

In the biological world, learning mechanisms are the result of evolutionary processes and, as summarized elsewhere (Floreano, Dürr & Mattiussi, 2008; Soltoggio, Stanley & Risi, 2017) many authors have studied the artificial evolution of machine learning, particularly in neural networks (e.g., Chalmers, 1990; Fontanari & Meir, 1991; Bengio, Bengio, Cloutier & Gecsei, 1992; Baxter, 1993; Dasdan & Oflazer, 1993; Nolfi & Parisi, 1996; Char, 1997; Kirchkamp, 1999; Federici, 2005; Niv, Joel, Meilijson & Ruppin, 2002; Di Paolo, 2003; Heller, 2004; Soltoggio, Dürr, Mattiussi & Floreano, 2007; Soltoggio, 2008; Soltoggio, Bullinaria, Mattiussi, Dürr & Floreano, 2008; Dunlap & Stephens, 2009; Soltoggio & Jones, 2009; Risi, Vanderbleek, Hughes & Stanley, 2009; Risi, Hughes & Stanley, 2010; Risi & Stanley, 2010; Risi & Stanley, 2012; Ellefsen, 2013; Ellefsen, 2014; Lehman & Miikkulainen, 2014; Orchard & Wang, 2016). However, much research remains to even determine appropriate environments in which to evolve artificial learning mechanisms. Evidence from the biological world can inform this research. “The ability to confront novel stimuli, learn about them, and adjust behavior is a hallmark of both intelligence and self-awareness. The evolution of intelligence is costly, in both the development of the neural tissue necessary to process the information and its metabolic maintenance. In addition, the evolution of intelligence requires dramatic changes in life history patterns, such as long juvenile phases and high parental investment per offspring. For most species, these costs, measured as reductions in reproductive fitness, far outweigh the costs of an occasional inappropriate use of [fixed action

patterns], and extensive intelligence has not evolved in many animal groups" (Campbell & Reece, 2008).

The observation that *nurturing*—the contribution of time, energy, or other resources by one individual to the expected physical, mental, social, or other development of another individual with which it has an ongoing relationship—might promote the evolution of learning in the biological world led us to the realization that a similar situation might exist in the artificial world (Woehrer et al., 2012), an idea which has been recently echoed by others (Soltoggio et al., 2017), and to begin studies of the artificial evolution of nurturing and learning (Leonce, Hoke & Hougen, 2012; Eskridge & Hougen, 2012). Moreover, these ideas lead us to posit the existence of a virtuous cycle between nurturing and learning in which each reinforces the other, leading to ever greater nurturing and ever greater learning in some lineages, as illustrated in Figure 1 (Woehrer et al., 2012).

Reinforcement learning is of particular interest to the research community (Dayan & Niv, 2008; Niv, 2009; Botvinick, Niv & Barto, 2009; Daw & Frank, 2009; Doll, Jacobs, Sanfey & Frank, 2009; van Hasselt, 2012; Shah, 2012; Wiering & van Otterlo, 2012; Kober, Bagnell & Peters, 2013; Kormushev, Calinon & Caldwell, 2013; Dridi & Lehmann, 2014; Ghanbari, Vaghei, Noorani & Reza, 2014; Kober & Peters, 2014b; Shteingart & Loewenstein, 2014; Pagliuca & Nolfi, 2015; Amarjyoti, 2017; Li, 2017; Polydoros & Nalpantidis, 2017) as is learning in changing environments (Plotkin & Odling-Smee, 1979; Stephens, 1991; Anderson, 1995; Krakauer & Rodríguez-Gironés, 1995; Kirchkamp, 1999; Heller, 2004; Nakahashi, 2007; Dunlap & Stephens, 2009; Kendal, Giraldeau & Laland, 2009; Dukas, 2013; Snell-Rood, 2013; Aoki & Feldman, 2014). Here we have chosen to look at those two concepts in conjunction with one another from an evolutionary standpoint.

Naturally, learning is likely to evolve in some environmental niches but not in others (Johnston, 1982; Kerr & Feldman, 2003). However, it might not be necessary for an evolving lineage of organisms to simply "hit on" such a niche—it may be possible for niches to be carved out by the organisms themselves (Kirsh, 1996; Kerr & Feldman, 2003) and nurturing may be a vital part of the niche carved out by organisms that subsequently evolve learning. Here we study whether a niche created by the addition of nurturing facilitates the evolution of learning within that niche.

The only alternative to learning considered in most early research on the artificial evolution of learning has been random behavior (e.g., Chalmers, 1990; Fontanari & Meir, 1991; Bengio et al., 1992; Dasdan & Oflazer, 1993; Char, 1997; Di Paolo, 2003; Federici, 2005; Soltoggio et al., 2007) and we have previously done likewise (Shah & Hougen, 2017a). However, research has increasingly (e.g., Baxter, 1993; Nolfi & Parisi, 1996; Kirchkamp, 1999; Niv et al., 2002; Soltoggio, 2008; Soltoggio et al., 2008; Dunlap & Stephens, 2009; Soltoggio & Jones, 2009; Risi et al., 2009; Risi et al., 2010; Risi & Stanley, 2010; Risi & Stanley, 2012; Ellefsen, 2013; Ellefsen, 2014; Lehman & Miikkulainen, 2014; Orchard & Wang, 2016) also considered the alternative of *instincts*—innate, typically fixed actions in response to particular stimuli. We find it more compelling to also consider instincts because random behavior is such a low baseline and because we know that many organisms that exhibit little learning do not simply act at random. For this reason, we here (and elsewhere, see Hoke, 2017) allow for the possibility of evolving instincts in addition to the possibilities of evolving learning or random behavior.

Finally, many types of nurturing are possible (Woehrer et al., 2012; Eskridge & Hougen, 2012). Here we have chosen to look at nurturing as task simplification. That is, the individual is nurtured by having a part of the task performed on its behalf so that it can focus its learning efforts on the portion of the task remaining. (Clearly, the nurturer must leave a learning opportunity for the nurtured individual within that remaining portion or there will be no benefit to the evolution of learning.)

2 Hypotheses

This study hypothesizes that nurturing promotes the evolution of learning. What this means is that in the nurturing niche, learning is more likely to be useful and therefore apparent than it is in the non-nurturing niche. If this hypothesis is true it could manifest itself in two primary ways: First, nurturing might increase the likelihood of evolving worthwhile learning. Secondly, learning evolved in

the nurturing niche might outperform learning evolved in the non-nurturing niche. This gives us two primary hypotheses:

H_1 Learning will evolve more frequently with nurturing than without.

H_2 Average performance of learning agents evolved in the nurturing niche will exceed that of learning agents evolved in the non-nurturing niche.

We can further think of this either categorically (with several possible categories of learning system performance) or in terms of reward received (a performance continuum) and also whether the individual's behavior is entirely learned or could be influenced by instincts. Considering the various possible combinations of each of these aspects of the hypotheses gives numerous possible sub-hypotheses, which will not be enumerated here for the sake of brevity[1]. These concepts are operationalized in Section 4.2.

3 Approach

The overall approach consists of an environment in which learning should be favored by evolution, a simple artificial neural network and associated parameterized learning rule suitable for a learning robot controller, and a genetic algorithm to carry out the artificial evolution.

3.1 Environment

To test the hypothesis that nurturing promotes the evolution of learning, we need an environment in which learning should be favored over other alternatives. After all, if an outcome is not favored, an evolutionary process cannot be expected to arrive at that outcome nor would arriving at that outcome be desirable.

There are, of course, environments that favor instincts over learning or random behavior, favor learning over instincts or random behavior, and favor neither learning nor instincts over random behavior (see, e.g., Dukas, 1998; Kerr & Feldman, 2003; Ellefsen, 2013).

For example, if an environment never changes, inherited instincts can define an optimal policy and, because the exploration required for reinforcement learning necessarily deviates from that optimal policy, this environment would favor instincts over learning.[2] Naturally, random behavior is vanishingly unlikely to match the optimal policy in such an environment. We call an environment in which instincts are favored over learned or random behaviors, an *instinct positive environment*.

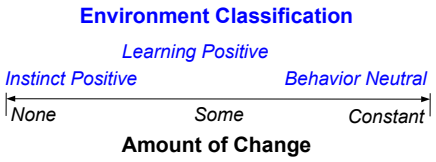


Figure 2. Conceptual diagram of the relationships between degree of environmental change and evolutionarily favored responses. As the degree of environmental change ranges from none to constant, the evolutionarily favored response shifts from instincts to learning to behavior neutral. Note that the region that favors instincts and learning overlaps.

At the other extreme, if the environment is constantly changing such that current experience does not help to predict future reward, then learning will provide no advantage, yet neither will instincts—the individual might just as well behave randomly. We call such an environment a *behavior neutral environment*.

Finally, if an environment changes infrequently during an individual's lifetime, its inherited instincts will not lead to maximum reward on the time steps after change, so learning might prove valuable provided that there is sufficient time to exploit the knowledge gained via learning. Again, random behavior is vanishingly unlikely to match the optimal policy either before or after the environmental changes. This gives something of a "sweet spot" between too little change (which favors instincts over learning) and too much change (which provides no advantage to instincts, learning, or random behavior). We call any environment in this sweet spot a *learning positive environment*.

These environmental classifications and their conceptual relationships to environmental change are shown in Figure 2. Note that the evolution of instincts is an adaptation that happens across generations whereas learning is an adaptation that happens within an individual's lifetime. However, these adaptations are not exclusive, so it is likely that there is also a rate of environmental change that would favor both instincts and learning and thus these categories are shown overlapped.[3]

The basic setup of the experiments is inspired by the light switching arena of Floreano & Urzelai (2000) while in implementation and experimental design it is a modification of the setup by Leonce et al. (2012) where at one end of the arena is the light source and at the other end is the light switch. An agent moves across the arena and its goal is to get to a light source in minimal time. In the original experiments of Floreano & Urzelai (2000), the robot needs to turn on the switch in order to collect energy from the light. In the experiments of Leonce et al. (2012), the light switch can be turned on by the robot itself, in which case the entire behavior is known as *self care*, or it can be turned on for the robot by a second robot that is also present in the arena, in which case the second robot is said to nurture the first robot. In the present paper, there is only one robot present in the arena for each trial as in Floreano & Urzelai (2000) but, inspired by Leonce et al. (2012), the light switch is either turned on for the robot prior to each trial to provide the nurturing treatment case or turned off at the beginning of each trial to provide the non-nurturing or self-care control case.

In addition to the above setup, the experiments require an additional component of an infrequently changing environment that is essential to provide a learning positive environment. Niv, et al. (2002) show in a bee foraging experiment that the evolution of learning in a terminal reward scenario can be accomplished using hebbian and antihebbian learning mechanisms. The three most important aspects of the bee foraging experiments that are not present in the previous light-switching experiments by Leonce et al. (2012) are (1) multiple possible targets (which are different colored flowers in the bee experiments) with different reward values, (2) the rewards of these targets change both between and within generations, so an individual needs to learn in order to perform well, and (3) each individual has multiple attempts (trials) at a terminal reward task, so each individual has an opportunity to learn during its lifetime. Thus, the experimental design in this research is a fusion and extension of these previous experiments.

Here there are three lights of different colors with different reward values: high (0.9), medium (0.5), and low (0.1), where, it should be noted, the medium value is halfway between the high and low values. The reward values change both between generations and during the lifetime of each individual, thus each individual has to learn in order to acquire a high level of reward from the environment. The setup consists of an individual robot that starts from the center of the arena and aims to find a path that will maximize its reward.

In the case of nurturing, the switch and thus the lights are already turned on for the robot. The robot is being nurtured externally. The robot starts each trial of its life looking for the best rewarding light source. In the case of self-care, the switch is turned off at the start of every trial. Thus the robot has to first travel to the switch and then look for the best rewarding light source.

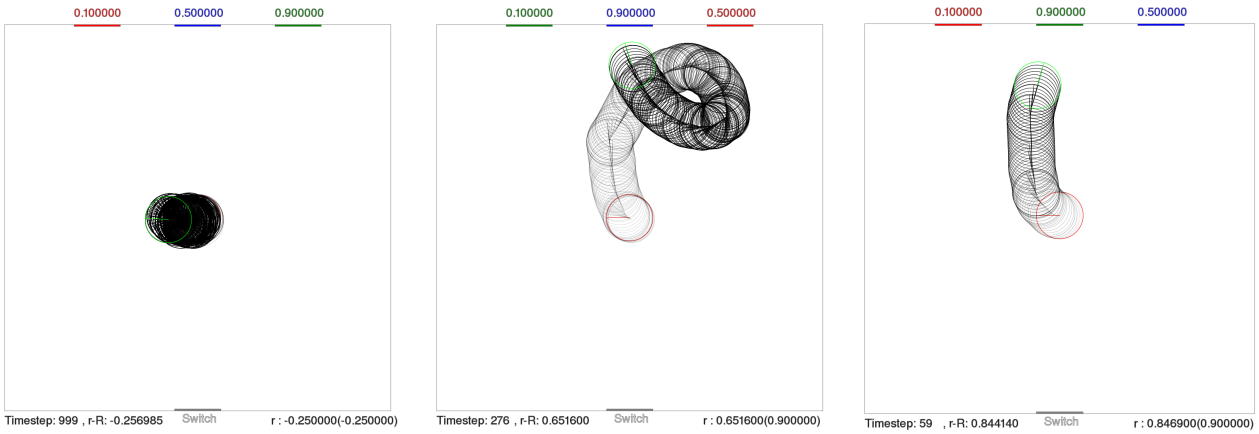
In both the nurturing and self-care cases, the individual has 1000 time steps within which to try to reach one of the lights while it is on. If the individual reaches a light that is on before the time expires, it receives a reward based on which light was reached and how quickly it reached that light (quicker is better), the trial ends, and the individual is returned to the starting location for the start of the next trial. The positive reward value for each trial is calculated using

$$r = \frac{T - t}{T} r_v, \tag{1}$$

where r stands for the scaled reward calculated, T is total number of time steps per trial (1000), t is the time step on which individual reaches the light source, and r_v is the raw reward value of the light reached (i.e., 0.9, 0.5, or 0.1). If the individual does not reach any light before the trial ends, it receives a penalty of -0.25 and is likewise returned to the starting location for the start of the next trial. This continues for a lifetime of 2000 trials.

Some examples of robots acting in this environment (under the nurturing condition) are shown in Figure 3.

Halfway through the lifetime of the robot, the highest reward value is swapped with the lowest reward value to change the environment and encourage learning. It is important to note that a successful instinctive individual with no learning capability visits the same light source over and over again, so the maximum reward it is able to collect in its lifetime is a moderate reward by following one of the following three strategies: (1) visiting a light that provides it a medium reward throughout its lifetime, (2) visiting a



(a) Poor performance. The robot spun in circles for the entire duration of the trial (1000 time steps).

(b) Intermediate performance. The robot reached the high rewarding light but the path contained an extraneous loop.

(c) Near-optimal performance. The robot moved to the high rewarding light at high speed and via a direct path.

Figure 3. Plots showing performance of the robot in the arena. The light gray outline shows the walls of the arena and the bars along the "north" wall colored red, green, and blue show the locations of the colored lights in the arena. Note that the colors and reward values of the lights change between and during lifetimes. The location of the switch is shown with the gray bar on the opposite wall and does not change. Circles show the robot's position while lines show its heading on each time step. The robot's initial position is represented by a red circle and its final position is represented by a green circle. All the steps in between are represented by lighter gray (earlier steps) to darker gray (later steps). Below each arena figure, "Timestep" shows the total number of time steps taken by the robot during that trial, "r-R" shows the current scaled reward minus the average reward and "r" shows the current scaled reward received and, parenthetically, the raw reward for the light reached. In (a), no light was reached so the reward value was -0.25, whereas for (b) and (c) the high rewarding light was reached in each case.

light that provides it a high reward during the first half of its lifetime but a low reward during the second half of its lifetime, or (3) visiting a light that provides a low reward during the first half of its lifetime but a high reward during the second half of its lifetime.

Having a difference of the nurturing (treatment) and non-nurturing (control) conditions while keeping everything else the same, the expectation is that the data collected will highlight the nurturing vs. self-care performance differences.

3.2 Artificial Neural Network (ANN)

In the experiments of Leonce et al. (2012), each robot is an Enki-based simulation (Magnenat, Waibel & Beyeler, 2007) of an e-puck robot (Mondada et al., 2009) controlled by a single-layer feed-forward artificial neural network (ANN) for which each input x_i is based on what the robot observes with its camera and each of the two outputs y_j controls the speed of one of the robot's two drive wheels. In those experiments, a genetic algorithm is used to evolve the single layer of weights $w_{i,j}$ of the ANN and there is no mechanism for the robot to adjust those weights during its lifetime. In other words, the robots' behaviors could be considered evolved instincts; they are not learned.

Here, however, we want to additionally allow for the weights to be learned and for the learning mechanism to be evolved. Following the seminal work of Chalmers (1990) as well as that of several others studying the evolution of learning (e.g., Fontanari & Meir, 1991; Baxter, 1993; Dasdan & Oflazer, 1993; Niv et al., 2002; Di Paolo, 2003; Soltoggio et al., 2007), we chose to provide the structure of an appropriate weight update rule and allow evolution to select the parameters for it. For the structure, we chose that of the Stochastic Synapse Reinforcement Learning (SSRL) algorithm (Shah & Hougen, 2017b). In SSRL, on each time step each ANN weight is sampled from a uniform distribution with a specified mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$ for that synapse. When reward is received by the system, SSRL uses eligibility traces to adjust the means and standard deviations of its weights. In this way, an SSRL system learns not only appropriate responses for given inputs (as given by the means) but also the degree to which it should be exploratory in its behaviors (as given by

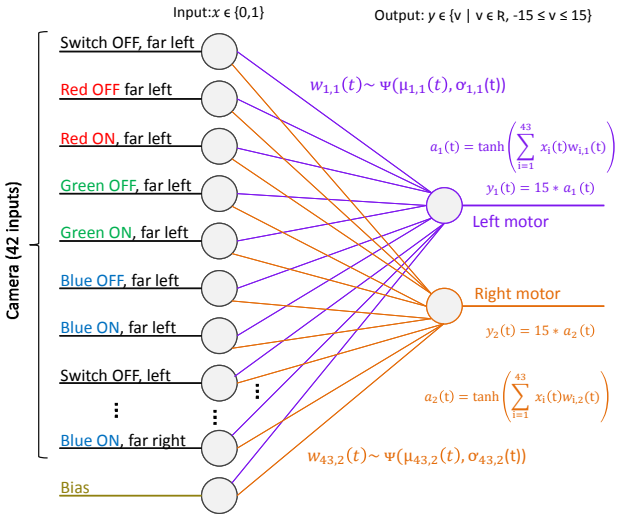


Figure 4. SSRL ANN Implementation. Fully connected feed-forward neural network with 42 input units, a bias unit, no hidden units, and two output units.

the standard deviations). In addition, SSRL uses a sliding window of past experiences to determine the collection of experiences over which expected reward is calculated. Our SSRL ANN is shown in Figure 4.

SSRL has eight parameters that can be specified by the system designer or, as in the present study, evolved. The SSRL parameters are described and symbols for them are given in Table 1, which also shows the calculations used to arrive at their values based on their genes, as explained in Section 3.3.

The inputs to the robot, and thus to the ANN, come from a linear color camera, 60 pixels wide, facing forward. In order to convert this input into something meaningful for the neural network, these 60 pixels are divided into 6 subgroups of 10 pixels each. These 6 subgroups represent far left, left, near left, near right, right, and far right regions of the camera's field of view. Inside each camera region, the robot looks for seven different color values, corresponding to seven different environmental features of interest—the color of the switch when it is turned off, the color of the red light when it is turned off, the color of the red light when it is turned on, the color of the green light when it is turned off, the color of the green light when it is turned on, the color of the blue light when it is turned off, and the color of the blue light when it is turned on. Note that the robot does not recognize the color of the switch when it is turned on because once the switch is turned on the robot does not need to pay attention to it for any particular action.

If 5 or more out of 10 of the color pixels for one of the seven colors are found to be present in a given camera region, then an input of 1 is given to the corresponding neuron in the neural network as shown in Figure 4. Conversely, if 4 or fewer of the pixels are of any particular color, an input of 0 is passed to the corresponding input neuron. Thus, considering six regions for each of the seven possibilities, there are a total of 42 binary sensory inputs to the neural network. In addition, a bias unit is used to input a constant 1 value to the ANN. This can be useful, for example, if the robot does not see any of the above mentioned objects (i.e., it is facing a blank wall) as it allows the robot to move in the arena even when all of its sensory inputs are 0. This makes the total count of binary input units 43 and thus the number of connection weights between the input and output units is 86, as this is a fully connected feed-forward neural network.

The binary input values may change at every time step as the robot moves around the arena. The camera input depends upon what is in its field of view. The robot is determined to have reached the light (thus ending the trial), when all six of its visual input regions register the presence of the same light source. Similarly, an individual is determined to have reached the switch and have turned it on when all six of its visual input regions register the presence of the switch. This is robot-centric.

The robot's output units consist of two neurons, each representing one of the robot's differential wheels. The activation value for each of the two output motor units is computed using the weighted sum of the inputs connected to that output unit squashed using a hyperbolic tangent function. The squashed real-valued activations are then scaled up to generate the output motor speed in the range -15 to $+15$ (to scale them to the minimum and maximum values found in the Enki simulator).

3.3 Genetic Algorithm (GA)

Next, it is essential to introduce an evolutionary process through which the learning algorithm parameters can be evolved. Further, an option is needed to evolve instincts together with learning parameters to demonstrate that

Table 1. Learning parameter descriptions, symbols, and their calculated scaled values. Here g_l is the gene at locus l.

Name	Symbol	Calculation
Mean (μ) Learning Rate	${}_{\mu}\eta$	${}_{\mu}\eta = g_0$
Standard Deviation (σ) Learning Rate	${}_{\sigma}\eta$	${}_{\sigma}\eta = g_1$
Minimum Standard Deviation	${}_{min}\sigma$	${}_{min}\sigma = 0.5g_2$
Maximum Standard Deviation	${}_{max}\sigma$	${}_{max}\sigma = 0.5g_3 + 0.5$
Initial Standard Deviation	${}_{init}\sigma$	${}_{init}\sigma = g_4({}_{max}\sigma - {}_{min}\sigma) + {}_{min}\sigma$
Sliding Window Size	s	$s = g_5T$ (where T is maximum time steps per trial)
Mean Decay Rate	${}_{\mu}d$	${}_{\mu}d = 10^{2g_6-1}$
Standard Deviation Decay Rate	${}_{\sigma}d$	${}_{\sigma}d = 10^{2g_7-1}$

nurturing promotes the evolution of learning both with and without the possibility of instincts. In the ANN control systems used in this work, instincts correspond to the initial mean values of the synapse weights, as these are the primary determinants of an individual's behavior unless and until they are adjusted based on experience. In the experiment in which only learning rule parameters may be evolved, the synapse weight mean values are randomly initialized. This means that an individual cannot inherit its instincts from its ancestors, which means that instincts cannot be evolved. In this experiment, only the learning rule parameters are encoded in each individual's chromosome, so only learning can be evolved. In contrast, for the experiment in which we want to allow learning and/or instincts to evolve, the initial mean values of each synapse weight are also encoded in each individual's chromosome.

We use a generational genetic algorithm in which fitness is defined to be the total reward collected in the arena by an individual during its lifetime. The GA works as follows:

1. Chromosomes for all individuals in the starting population (generation 0) are randomly initialized.
2. All individuals are evaluated independently.
3. After all individuals are evaluated, selection is performed to determine the composition of the next generation.
   (a) First, zero or more individuals are copied without changes to the next generation in order of fitness. These unaltered copies of the most fit individuals are known as *elites*.
   (b) Next, *clones* are added to the new generation. Clones differ from elites in that clones are not necessarily the most fit individuals from the population and they may undergo mutation. For each clone, a tournament bracket of size b is formed and b individuals are selected at random (with replacement) from the population to fill it. The individual in the tournament with the highest fitness is selected as the winner. The winner is cloned, possibly with mutation, and the clone is placed into the new generation. This process is repeated until the desired number of clones has been added to the new generation.
   (c) Finally, non-clonal offspring are added to the generation. These offspring are generated by performing two tournaments to find two winners and then using uniform crossover on the two winners to produce two offspring. Each offspring then has a chance of undergoing mutation. The process repeats until the size limit of the new population is reached. Note that the same individual can win multiple tournaments, thus it can crossover with itself to generate two offspring.
4. During mutation, there is a small chance that a given gene will be mutated. If selected for mutation, a normal distribution with zero mean is used to select the value to be added to the mutated gene. If the mutation would result in an allele outside the gene range limits (if any), the allele is set to be equal to nearest limit value.
5. The algorithm runs for a fixed number of generations.

Various parameters chosen for the genetic algorithm are listed in Table 2. Ten generations were determined to be sufficient for the evolutionary courses to diverge in the two niches (nurturing and self-care).

The eight learning rule parameters are encoded in each chromosome. Each parameter, a gene in the chromosome,

Table 2. Genetic Algorithm Parameters used and their Descriptions/Values.

GA Parameter	Description/Value
Population Size	30
Number of Generations	10
Chromosome Length	8 (learning) + 86 (instincts)
Selection Method	Tournament w/ Replacement
Tournament Size	3
Crossover Type	Uniform
Crossover Percentage	73%
Reproduction Method	1 Elite, 7 Clones, 22 Crossed
Gene Mutation Rate	5% per Gene
Gene Mutation StDev	0.1
Mutation Method	Normal Distribution

is a randomly generated value between 0 and 1 (inclusive). However, before the beginning of an individual's lifetime, scaling is required for some learning parameters to make them algorithmically plausible in the context of learning. The scaling details are shown in Table 1.

Learning rates ${}_{\mu}\eta$ and ${}_{\sigma}\eta$ do not need any scaling as a number between 0 and 1 is a valid learning rate for both μ and σ. Minimum standard deviation ${}_{min}\sigma$ is scaled to be in the range [0, 0.5] and maximum standard deviation ${}_{max}\sigma$ is scaled to be in the range [0.5, 1]. This ensures that the minimum exploration rate is in the lower half of the range of possible exploration rates while the maximum exploration rate is in the upper half of possible exploration rates. Initial standard deviation ${}_{init}\sigma$ is scaled to make sure that it is between the minimum and maximum scaled sigma values. Sliding window size s is multiplied by the trial size to ensure that the minimum size of the sliding window is zero and the maximum size is the length of an entire episode. The final two parameters are decay rates for μ and σ. Both of these parameters are scaled the same way and are described using

$$d = 10^{2g_l - 1}, \tag{2}$$

where g_l is the appropriate gene with a value sampled from a uniform distribution in [0, 1]. (This scales the value of d to [0.1, 10].). A normalization factor ν is derived from d using

$$\nu = 1/(d^0 + d^1 + d^2 + ... + d^{T-1}), \tag{3}$$

where T is the maximum time steps in a trial (The normalization factor is the sum of a geometric series).

The value of d in the above set of equations is the value that is considered the decay rate and is used to calculate eligibility values at all the time steps. To apply the normalization factor ν, assuming that the above calculations are performed for ${}_{\mu}d$ for the sake of example, then the normalization factor can be applied as

$$\Delta\mu_{ij}(\tau) = {}_{\mu}\eta\,(r(\tau) - \overline{r}(\tau)) \sum_{k=1}^{t} {}_{\mu}e_{ij}(k)\,{}_{\mu}d^{(t-k)}\nu \tag{4}$$

where $\Delta\mu_{ij}(\tau)$ is the change in μ (mean) on link i, j on trial number τ, $r(\tau)$ is the reward received during that trial, $\overline{r}(\tau)$ is the expected reward during that trial estimated using the sliding window, ${}_{\mu}e_{ij}(k)$ is the eligibility at time step k, and t is the total time steps in this trial.

In Equation 4, the normalization factor ν keeps the total of the discount factors applied to the eligibilities at less than or equal to one. A gene value g used in the above equations will function as follows:

- $g < 0.5$ means give more importance to the recent actions in this trial,
- $g = 0.5$ means give equal importance to all the actions in this trial,
- $g > 0.5$ means give more importance to the earlier actions in this trial.

To allow for the evolution of instincts, initial weights of the ANN are added to the chromosome along with the learning rule parameters. As the initial weights are passed from the parent population's successful individuals to the offspring with little or no change, they can be considered instincts. The objective is not only to answer hypotheses related to the evolution of learning and instincts but also to see if the main hypothesis still holds true after letting instincts evolve together with learning. This would help to indicate the generality of this approach.

In the experiment where only learning is evolved (no instincts), the nurtured individual only has to learn one thing, i.e., to go to the high-rewarding light source whereas the non-nurtured individual has to learn two things, i.e., to go to the switch and then to the high-rewarding light. However, in the experiment in which the evolution of both learning and instincts are allowed, both nurtured and non-nurtured individuals only need to learn one thing, i.e., to go

to the high-rewarding light. This is because nothing changes about the switch either within or between lifetimes, i.e., switch position and behavior are constants. This means that a lineage could evolve instincts to turn on the switch and then individuals in that lineage would only need to learn about the lights. Allowing individuals to evolve instincts (for the non-changing parts of the environment) should aid in the evolution of learning in the non-nurtured niche. Nonetheless, the nurtured niche is still distinct from the non-nurtured niche. In the nurtured niche, the individual only needs to carry out one action, whereas in the non-nurtured niche the individual needs to carry out two actions. This still provides an advantage to individuals in the nurtured niche and therefore we expect useful learning to appear more often and to a greater degree in the nurtured niche.

4 Experiments

In the present study, we have a single main hypothesis that can be manifested in one of two ways (category likelihood and performance continuum), may be influenced by the absence or presence of instincts, and may be looked for in subsets of the data as well as at the aggregate level.

This section explains the experiments carried out to test these hypotheses, the evaluation criteria used for categorical comparisons, and the method of data scaling used to ensure a fair comparison across conditions for continuous data.

4.1 Experimental Setup

Two experiments were performed. In *Experiment 1*, only learning may be evolved (the chromosome contains only the eight parameters of the SSRL algorithm). In *Experiment 2*, both learning and instincts may be evolved (the chromosome contains the eight parameters of the SSRL algorithm plus 86 initial weights for the ANN). Within each experiment, 30 repetitions of each condition (nurturing and self care) were performed to give sufficient data for meaningful statistical analysis. Note that both experiments test both hypotheses as both hypotheses concern the nurturing vs. self-care niches and both experiments contain both of these conditions. This also allows the results of these experiments to be combined and contrasted to determine trends across and between experiments.

4.2 Categorical Comparisons

To evaluate the data with respect to the first hypothesis (learning will evolve more frequently with nurturing than without), we need an objective way to determine whether or not learning has occurred. Simple adjustments to the mean values of the weights should not be considered sufficient evidence of learning, as learning should indicate that the adjustments improve the learner's performance. Moreover, we're not really concerned about minor improvements—we would like to see substantial improvements in behavior due to learning.

The boundary line we draw between substantial and non-substantial learning is based on the performance possible with instinctive behaviors alone. Any learning that performs better than the best theoretically possible instinctive behavior on average at the end of an agent's lifetime falls under the category of *substantial learning*. In contrast, performance that is lower than or equal to that of the best theoretically possible instinctive performance will be called *non-substantial*. Further, we divide the category of substantial learning into *good* and *moderate* learning, for those individuals who outperform the best theoretically possible instinctive individual in both halves of their lifetimes and those substantial learning individuals who outperform the best theoretically possible instinctive individual in exactly one half of their lifetimes, respectively.

To operationalize these terms with respect to these experiments, recall that the reward value for each trial is calculated using Equation 1. A search is performed for the lowest final time step value of any individual in the arena for both the nurturing and self-care niches. These values represent good approximations of the minimum amount of time in which an individual can complete the task(s) in each environment. That number is taken and 10% is added to that value to consider the possibility that the highest rewarding light might be located in the farthest corner[4] of the arena in order to calculate a fair value for each niche.

The best instinctive individual that always goes to the same light in both halves of its life should achieve a maximum of 0.47 on average in the nurturing niche and 0.41 in the self-care niche. Note that this is true whether

the instinctive individual goes to the non-changing, medium-rewarding light throughout its lifetime or goes to a light that switches rewards at the halfway point of the individual's life such that the individual receives the high reward in one half of its lifetime and the low reward in the other half. Thus any individual that gains a fitness higher than the cutoff value for its corresponding niche belongs in the substantial learning category while an individual with lower or equal fitness has performance that is poorer than or equal to the theoretical best instinctive performance and thus belongs in the non-substantial category. Moreover, a substantial learning individual that gains a fitness higher than the corresponding substantial value for both halves of its lifetime belongs in the good learning subcategory, whereas a substantial learning individual that exceeds the substantial learning value in only one half of its lifetime (and overall) belongs in the moderate subcategory.

4.3 Continuous Comparisons

In order to fairly compare the data between the nurturing niche and the self-care niche with respect to the second hypothesis (average performance of learning agents evolved in the nurturing niche will exceed that of learning agents evolved in the non-nurturing niche), it is important to have continuous performance data on the same scale.

While the base reward values of the lights are the same in both niches (0.9, 0.5, and 0.1), the fact that the optimal route to each light in the self-care niche is longer than the corresponding optimal route in the nurturing niche means that the best possible earned reward for each light is lower in the self-care niche and therefore normalization of earned reward is necessary.

A score termed *relative success* is calculated for each repetition. The *relative success* is a measure of how close the best individual in the final generation of that repetition is to the theoretical best omniscient individual in that niche. Note that this is not the same as the theoretical best instinctive individual, which goes to the same light every trial and receives (on average) the reward for moving quickly to the moderate rewarding light. Instead, this theoretical best omniscient individual moves quickly to the high rewarding light on every trial (or to the switch and then to the high rewarding light for the self-care niche), regardless of which light gives which reward, and does not need to spend time exploring. This relative success is compared between all repetitions for each niche as well as within the learning categories of each niche.

5 Results

Results are presented for each experiment individually and in aggregate. In addition, we consider differences found between Experiment 1 and Experiment 2.

5.1 Experiment 1: Evolution of Learning

Numerical results are presented along with exemplars showing the performance of individuals within both the nurturing and self-care niches. Numerical results are shown in Table 3.

Looking first at the categorical results, of the 30 repetitions of the nurturing condition, 29 individuals were evaluated to be substantial learners, leaving 1 non-substantial learner. The 29 substantial learners were further broken down in 27 good learners and 2 moderate learners. This contrasts with only 19 of 30 substantial learners for the self-care niche, leaving 11 non-substantial learners in that niche. The 19 substantial learners in the self-care niche are broken down into 8 good learners and 11 moderate learners.

These results can be statistically compared using Fisher's exact tests (two-tailed). Doing so, we find that counts are statistically significantly different in all cases considered, which are substantial versus non-substantial ($p = 0.0025$), good versus combined moderate plus non-substantial ($p < 0.0001$), and good versus moderate versus non-substantial ($p < 0.0001$).

Looking next at performance continuum results, we can see that the mean relative success values of the individuals evolved in the nurturing niche are higher than those of the individuals evolved in the self-care niche, both overall and within every category.

These results can be statistically compared using *t*-tests. Doing so, we find that the relative success scores are statistically significantly different both overall and in the substantial and good categories. However, there are too few moderate and non-substantial learners evolved in

Table 3. Results for Experiment 1: Evolution of learning. *Mean* is the arithmetic mean of the relative success values of individuals in the given category, *SD* is the standard deviation of the relative success values of those individuals, and *N* is the count of individuals within each category. Italics highlight statistically significant results.

Learner Category		Overall			Substantial			Good			Moderate			Non-Substantial		
		Mean	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
Niche	Nurturing	*81.6*	9.5	30	*82.5*	8.2	*29*	*83.6*	7.3	*27*	67.3	1.5	*2*	54.7	–	*1*
	Self-Care	*60.3*	9.1	30	*65.7*	6.9	*19*	*69.8*	7.4	*8*	62.7	4.7	*11*	51.1	3.0	*11*

the nurturing niche to say that the differences found are statistically significant.

To give some impression of what performance means in the various categories, we give examples of reward patterns for the most fit individuals from the final generation in the nurturing and self-care niches.

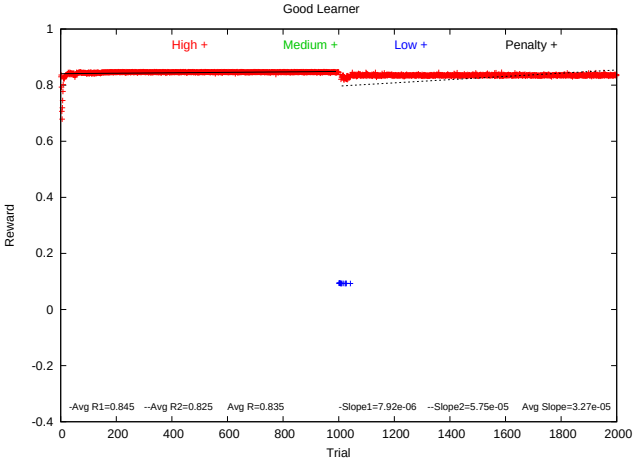


Figure 5. Evolution of Learning—Good (Substantial) Learning typical example in nurturing niche.

Figure 5 represents a typical good learning individual from the nurturing niche.[5] The first light encountered by this individual is the high-rewarding light and it learns within a few trials to move very quickly to that light. When that light becomes the low-rewarding light at Trial 1000, the individual receives a lower than expected reward for a few trials, then tries a different light that turns out to be the current high-rewarding light and quickly learns to prefer that light, moving to it quickly on each subsequent trial.

Figure 6 shows a moderate learning case from the nurturing niche, where the individual exhibits learning in both halves of its lifetime. However, in the second half of its lifetime it learns more slowly and therefore its average reward in the second half of its lifetime stays below the

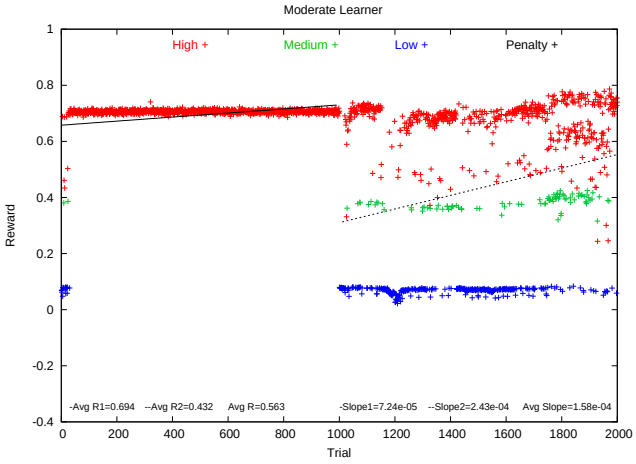


Figure 6. Evolution of Learning—Moderate (Substantial) Learning example in nurturing niche.

substantial-learning threshold, thus it is categorized as a moderate learner.

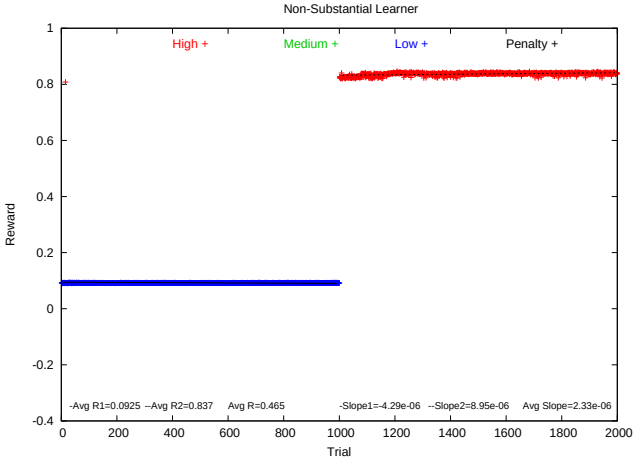


Figure 7. Evolution of Learning—Non-Substantial Learning example in nurturing niche.

Figure 7 shows the single non-substantial learner found among the results in the nurturing niche. It uses the initial weights that it was randomly assigned at the start of its

lifetime and never improves on them. It exploits the low-rewarding light during the first half of its lifetime and keeps going to that same light even after the change in the environment half way through the lifetime. It is interesting to note that it does see the highest rewarding light once early in its lifetime and finds quite an excellent path to it but never shifts its policy toward those actions. It is likewise interesting to note that this individual's overall behavior and fitness scores are very close to those of the theoretical best instinctive individual.

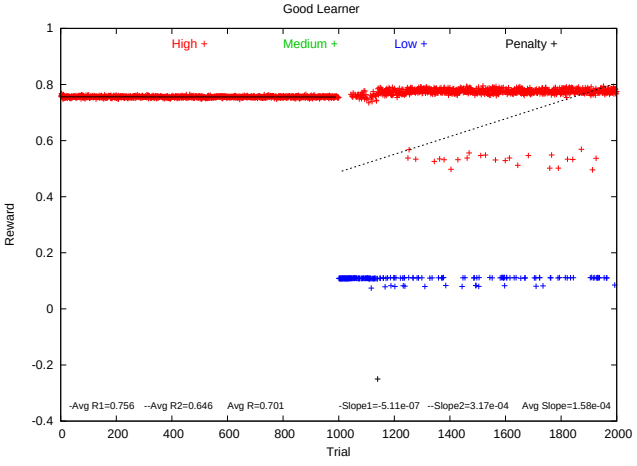


Figure 8. Evolution of Learning—Good (Substantial) Learning typical example in self-care niche.

Figure 8 is an example of a typical good learner from the self-care niche which happens to have good initial weights for going to the high-rewarding light source although it does not follow the optimal path. In the second half of its lifetime when the high-rewarding light it had been visiting becomes the low-rewarding light, it clearly shifts from that light to the new high-rewarding light and mostly exploits that resource with an approximately equally rewarding path as the one used in the first half of its lifetime.

Figure 9 shows an example of a typical moderate learner from the self-care niche that happens to have good initial weights to start with and thus never explores in the first half of its lifetime. In the second half of its lifetime, though, when the high-rewarding light that it had been targeting becomes the low-rewarding light, it somewhat shifts its focus from that light to a medium-rewarding light even though it explores the best light source on three trials.

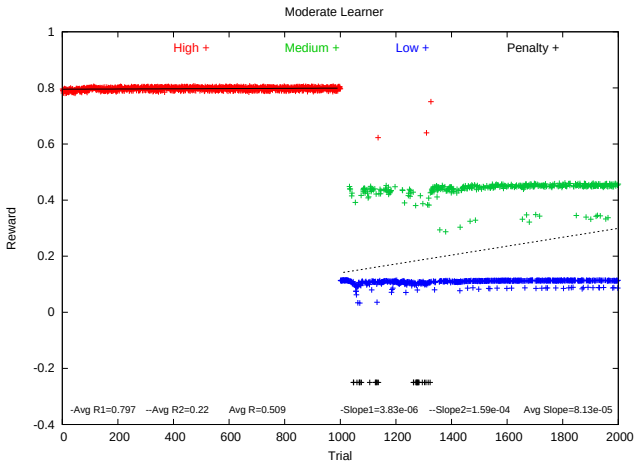


Figure 9. Evolution of Learning—Moderate (Substantial) Learning typical example in self-care niche.

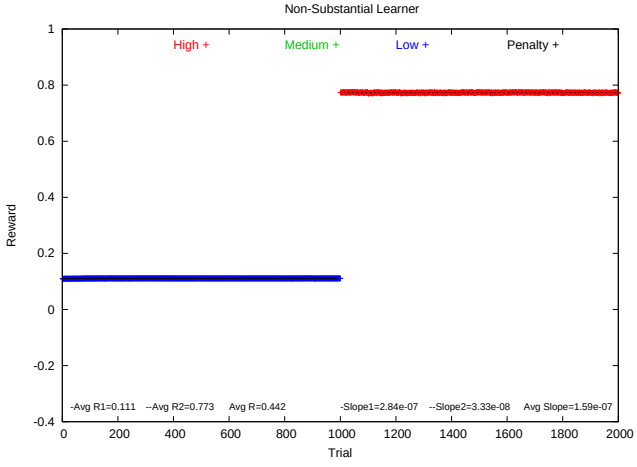


Figure 10. Evolution of Learning—Non-Substantial Learning example in self-care niche.

Figure 10 shows an example individual from the self-care niche that does not seem to learn anything useful throughout the course of its lifetime—it does not deviate substantially from the behaviors corresponding to its initial (random) weight means. Note that the individual depicted by Figure 10 is quite similar to the one shown by Figure 7 as neither individual appears to learn.

5.2 Experiment 2: Evolution of Learning and Instincts

As with the evolution of learning experiment, numerical results are presented along with exemplars showing the

performance of individuals within both the nurturing and self-care niches. Numerical results are shown in Table 4.

Again looking first at the categorical results, of the 30 repetitions of the nurturing condition, all 30 individuals were evaluated to be substantial learners, leaving no non-substantial learners. The 30 substantial learners were further broken down in 26 good learners and 4 moderate learners. This contrasts with only 24 of 30 substantial learners for the self-care niche, leaving 6 non-substantial learners in that niche. The 24 substantial learners in the self-care niche are broken down into 7 good learners and 17 moderate learners.

These results can likewise be statistically compared using Fisher's exact tests (two-tailed). Doing so, we find that counts are statistically significantly different in all cases considered, which are substantial versus non-substantial ($p = 0.0237$), good versus combined moderate plus non-substantial ($p < 0.0001$), and good versus moderate versus non-substantial ($p < 0.0001$).

Looking next at performance continuum results, we can see that the mean relative success values of the individuals evolved in the nurturing niche are higher than those of the individuals evolved in the self-care niche, both overall and within every category (except the non-substantial category, as none of the individuals from the nurturing niche tested poorly enough to fall into that category).

These results can be statistically compared using *t*-tests. Doing so, we find that the relative success scores are statistically significantly different both overall and in the substantial and good categories. However, there are very few moderate and no non-substantial learners evolved in the nurturing niche, so we cannot say that the differences found are statistically significant.

To again give some impression of what performance means in the various categories, we give examples of reward patterns for the most fit individuals from the final generation in the nurturing and self-care niches.

Figure 11 shows a typical good learning individual from the nurturing niche. The first light encountered by this individual is the high-rewarding light. Although it explores slightly different paths, it learns to move back to its original path. When that light becomes the low-rewarding light at Trial 1000, the individual receives a lower than expected reward for a few trials, then tries a different light that turns out to be the current high-rewarding light and quickly learns to prefer that light, moving to it quickly on each subsequent trial.

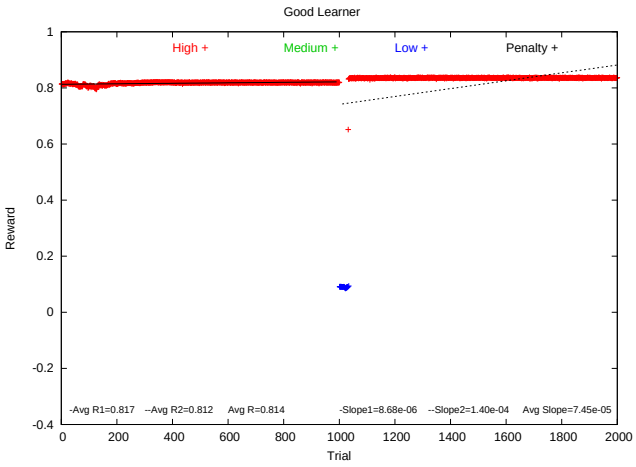


Figure 11. Evolution of Learning and Instincts—Good (Substantial) Learning typical example in nurturing niche.

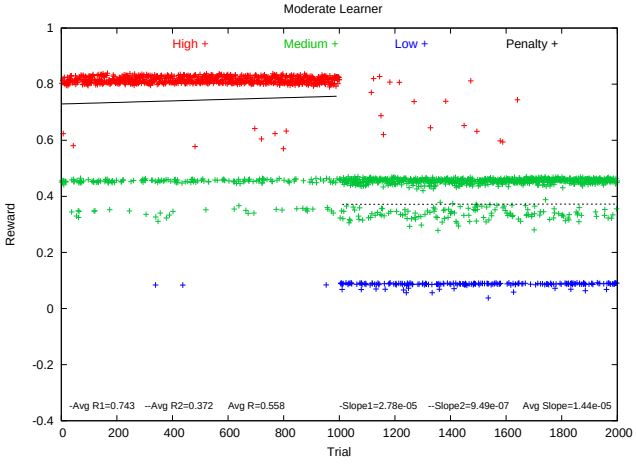


Figure 12. Evolution of Learning and Instincts—Moderate (Substantial) Learning example in nurturing niche.

Figure 12 shows a moderate learning individual from the nurturing niche that, during the first half of its lifetime, most frequently moves to the high-rewarding light but also often moves to the medium-rewarding light and rarely moves to the low-rewarding light. Also note that during this half of its lifetime, this individual goes to the high-rewarding light with increasing frequency. However, in the second half of its lifetime, when the light it had been seeking becomes the low-rewarding light, it mostly exploits its knowledge of the medium-rewarding light and mostly selects that light in

Table 4. Results for Experiment 2: Evolution of learning and instincts. *Mean* is the arithmetic mean of the relative success values of individuals in the given category, *SD* is the standard deviation of the relative success values of those individuals, and *N* is the count of individuals within each category. Italics highlight statistically significant results.

Learner Category		Overall			Substantial			Good			Moderate			Non-Substantial		
		Mean	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
Niche	Nurturing	*83.7*	8.7	30	*83.7*	8.7	*30*	*86.0*	6.9	*26*	69.1	2.6	*4*	–	–	*0*
	Self-Care	*65.2*	9.4	30	*67.9*	8.5	*24*	*78.2*	5.3	*7*	63.7	5.3	*17*	54.1	0.9	*6*

preference to the low-rewarding light. This behavior shift away from the light it had previously visited most frequently and to the medium-rewarding light is adaptive. However, it is not nearly as rewarding as if it had shifted to the new high-rewarding light (which it did find several times as it explored), so it does not collect enough reward on average during the second half of its lifetime to be categorized as a good learner. This individual is thus categorized as a moderate learner.

There is no non-substantial learning case found in the nurturing niche.

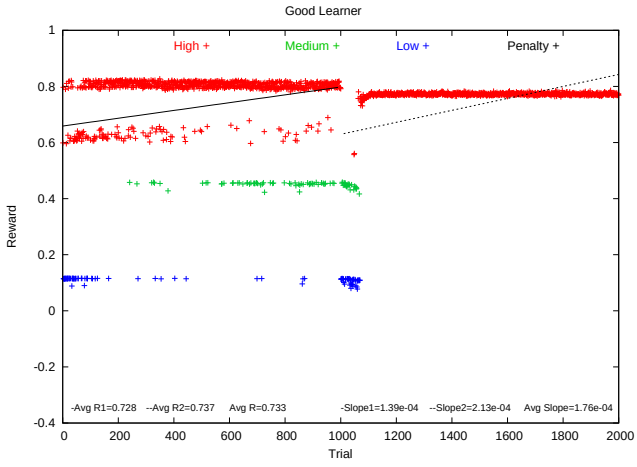


Figure 13. Evolution of Learning and Instincts—Good (Substantial) Learning example in self-care niche.

Figure 13 shows a good learning individual from the self-care niche that, during the first half of its lifetime, explores two different paths to the high-rewarding light and learns to prefer the quicker (higher rewarding) path. It also occasionally explores (initially) the low-rewarding light and (later) the medium-rewarding light. In the second half of its lifetime, once the light it was visiting becomes a low-rewarding light, it quickly finds a path to the best rewarding light. Although the paths followed are noticeably non-optimal, it still performs sufficiently well to be categorized as a good learner.

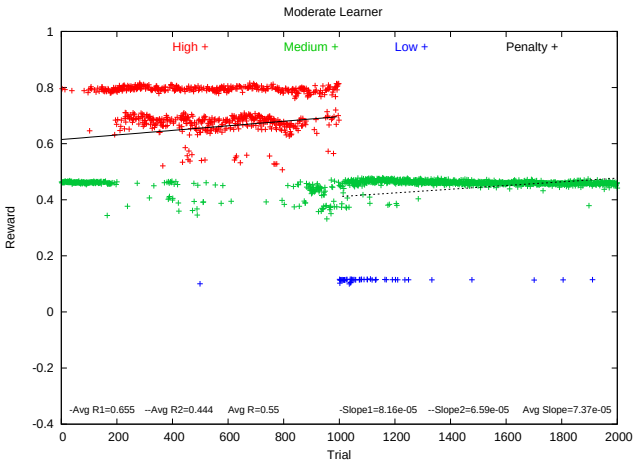


Figure 14. Evolution of Learning and Instincts—Moderate (Substantial) Learning example in self-care niche.

Figure 14 shows a moderate learning individual from the self-care niche that visits the medium- and high-rewarding lights during the first half of its lifetime and follows multiple paths to these lights. It learns during this half of its lifetime at least in part by shifting the frequency with which it executes each behavior. In the second half of its lifetime, when the high-rewarding light it was visiting becomes the low-rewarding light, it shifts its focus almost exclusively to the medium-rewarding light.

Figure 15 shows a typical non-substantial learner from the self-care niche. This individual starts by instinctively visiting the low-rewarding light and never improves on that. It also visits the medium-rewarding light occasionally; however, it never shows any learning to shift its focus to a better reward. When the rewards change at the midpoint of this individual's lifetime, the low-rewarding light it had been favoring becomes the high-rewarding light. However it continues to visit the medium-rewarding light occasionally.

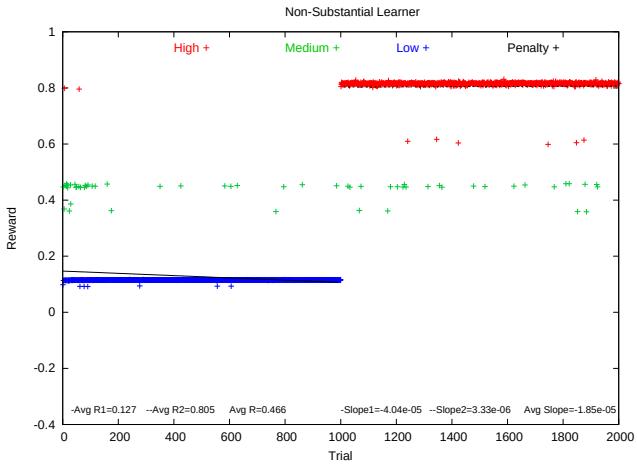


Figure 15. Evolution of Learning and Instincts—
Non-Substantial Learning typical example in self-care niche.

5.3 Evolution of Learning Regardless of Instincts

Combining the results from the previous two experiments, can give us further insight into the relationships between nurturing, learning, and instincts.

First, we combined data within niches across experiments, as given in Table 5. This aggregated data allows us to answer the more general question of whether nurturing promotes the evolution of reinforcement learning regardless of whether or not instincts are allowed to evolve.

We find that of the 60 total repetitions of the nurturing condition, 59 resulted in substantial learners and only 1 resulted in a non-substantial learner. The 59 substantial learners were broken down into 53 good learners and 6 moderate learners. In the self-care condition, the 60 total repetitions resulted in 43 substantial learners and 17 non-substantial learners. The 43 substantial learners were further categorized as 15 good learners and 28 moderate learners.

Comparing these results using Fishers exact tests (two-tailed), we find the counts are statistically significantly different in all cases considered, which are substantial versus non-substantial ($p = 0.0001$), good versus combined moderate plus non-substantial ($p < 0.0001$), and good versus moderate versus non-substantial ($p < 0.0001$). These results are entirely expected given the individual experiment results, of course, as the individual experiment comparisons were all congruent and statistically significant, meaning that combining them would simply result in smaller p values.

Looking at the combined performance continuum results, we can see that the mean relative success values of the individuals evolved in the nurturing niche are higher than those of the individuals evolved in the self-care niche, both overall and within every category. Again, this is entirely expected, given that this was true of each experiment independently.

Statistically comparing the combined results using *t*-tests, we find that the relative success scores are statistically significantly different both overall and in the substantial and good categories, as with the individual experiments. In addition, there are now enough data items in the moderate category to say that the higher mean relative success value there for the nurturing niche is also statistically significant. However, there is still just one non-substantial learner evolved in the nurturing niche, so no statistical conclusion can be reached about its apparent higher performance in that category.

Second, we can consider the differences between the results across experiments. In particular, it has been found elsewhere (Hoke, 2017) that allowing the genome to encode instincts can inhibit the evolution of learning. So, we might find that, in the experiment in which instincts can be encoded in the genome, learning evolves less frequently and/or results in lower performance than in the experiment in which instincts cannot be encoded in the genome. Alternately, because instincts and learning are not mutually exclusive but potentially complementary in this experiment (and, often, in nature), it is possible that we might find that learning evolves more frequently and/or results in higher performance in Experiment 2 than it did in Experiment 1. Indeed, to put a finer point on it, because it is possible for an individual to evolve both instincts and learning, because instincts and learning can combine to provide higher performance, and because we here measure learning based on performance, there is a reasonable possibility that we'll find more and/or better learning in Experiment 2 than in Experiment 1. These differences are shown in Table 6.

Looking at the categorical results, we can see that in the nurturing niche there were 30 substantial learners evolved when instincts were possible as compared to 29 when

Table 5. Results for combined experiments. *Mean* is the arithmetic mean of the relative success values of individuals in the given category, *SD* is the standard deviation of the relative success values of those individuals, and *N* is the count of individuals within each category. Italics highlight statistically significant results.

Learner Category		Overall			Substantial			Good			Moderate			Non-Substantial		
		Mean	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
Niche	Nurturing	*82.7*	9.1	60	*83.1*	8.4	*59*	*84.8*	7.1	*53*	*68.5*	2.3	*6*	54.7	–	*1*
	Self-Care	*62.8*	9.5	60	*66.9*	7.8	*43*	*73.7*	7.6	*15*	*63.3*	5.0	*28*	52.2	2.9	*17*

Table 6. Results for differences between experiments. Δ*Mean* is the difference in arithmetic mean between Experiment 1 and Experiment 2 of the relative success values of individuals in the given category and ΔN is the difference in count between Experiment 1 and Experiment 2 of individuals within each category. None of the results are highlighted in italics as none of the results are considered statistically significant.

Learner Category		Overall		Substantial		Good		Moderate		Non-Substantial	
		Δ*Mean*	ΔN	Δ*Mean*	ΔN	Δ*Mean*	ΔN	Δ*Mean*	ΔN	Δ*Mean*	ΔN
Niche	Nurturing	-2.1	0	-1.2	-1	-3.6	$+1$	-1.8	-2	–	$+1$
	Self-Care	-4.9	0	-2.2	-5	-10.4	$+1$	-1.0	-6	-3.0	$+5$

instincts were not possible, leaving no non-substantial learners with possible instincts and 1 non-substantial learner without the possibility of instincts. Looking within the substantial learners, of 30 where instincts are possible, 26 were good and 4 were moderate. This compares to the 29 substantial learners without the possibility of instincts, of which 27 were good and 2 were moderate. This means that when the possibility of instincts was added, the number of substantial learners decreased by 1 while the number of good learners increased by 1, the number of moderate learners decreased by 2, and the number of non-substantial learners decreased by 1. These differences do not approach statistical significance (with p values from Fisher's exact test, two tailed, ranging from 0.55 to 1.0).

In the self-care niche, the categorical results show 24 substantial learners evolved when instincts were possible as compared to 19 when instincts were not possible, leaving 6 non-substantial learners with possible instincts and 11 non-substantial learners without the possibility of instincts. Looking within the substantial learners, of 24 where instincts are possible, 7 were good and 17 were moderate. This compares to the 19 substantial learners without the possibility of instincts, of which 8 were good and 11 were moderate. This means that when the possibility of instincts was added, the number of substantial learners decreased by 5 while the number of good learners increased by 1, the number of moderate learners decreased by 6, and the number of non-substantial learners increased by 1. These differences also do not approach statistical significance (with p values from Fisher's exact test, two tailed, ranging from 0.24 to 1.0).

If these differences are combined between the nurturing and self-care niches, we see 54 substantial learners evolved when instincts are possible as compared to 48 when instincts are not possible, leaving 6 non-substantial learners with possible instincts and 12 non-substantial learners without the possibility of instincts. Looking within the substantial learners, of 54 where instincts were possible, 33 were good and 21 were moderate. This compares to the 48 substantial learners without the possibility of instincts, of which 35 were good and 13 were moderate. This means that when the possibility of instincts was added, the number of substantial learners decreased by 6 while the number of good learners increased by 2, the number of moderate learners decreased by 8, and the number of non-substantial learners increased by 6. These combined differences likewise do not approach statistical significance (with p values from Fisher's exact test, two tailed, ranging from 0.16 to 0.85).

Looking at the difference in performance continuum results, we see that the mean relative success values of the individuals evolved with the possibility of instincts were higher overall and in all categories than those of their counterparts evolved without the possibility of instincts, resulting in the negative difference values shown in Table 6. The differences ranged from -1.0 for the moderate learners

in the self-care niche to -10.4 for the good learners, also in the self-care niche.

Comparing the performance continuum results for the cases where instincts are not possible to those where they are, two of these comparisons would be considered statistically significant at an α value of 0.05 if the tests were considered in isolation. (These two are for good learners in the self-care niche and overall in that niche, with p-values of 0.028 and 0.049, respectively.) However, because these tests are not independent and we wish to maintain a family-wide error rate of 5%, the p-values needed for statistical significance were adjusted using the Hold-Šidák method, which resulted in none of the p-values being considered significant.[6]

6 Discussion

Both category likelihood and performance continuum results indicate that the nurturing niche overwhelmingly favors the evolution of learning as compared to the self-care niche.

All categorical comparisons between the nurturing and self-care niches for both Experiment 1 and Experiment 2 and their combined results are congruent and are found to be statistically significant.

In the performance continuum results, the mean performance of individuals from the nurturing niche is consistently higher than that for individuals from the self-care niche for both Experiment 1 and Experiment 2. Almost all of these results are statistically significant, except those few cases at the bottom of the performance spectrum where the individuals from the nurturing niche were too few to make solid statistical statements. This is expected as nurturing overwhelmingly outperforms self-care in terms of evolving good learners, so very few non-substantial or even moderate learners are found in the nurturing niche. A small number of samples means that a *t*-test will lack power, so it isn't surprising that the null hypothesis cannot be rejected in these cases, even though the means appear to be higher for the nurturing condition for both hypotheses. While no conclusions should be drawn from the results for which too few samples are present, they appear to be congruent with the results that are statistically significant and do not detract from the overall conclusion.

Altogether, the congruency of the results found and the statistical comparisons of these results very strongly indicate that both hypotheses and the main hypothesis from which they are derived are supported—that nurturing promotes the evolution of reinforcement learning in changing environments.

Considering the question of how the possibility of evolving instincts might influence the evolution of learning, and looking first at the categorical difference data, we find that in both the nurturing and the self-care condition, there are more substantial learners (and fewer non-substantial learners) when instincts can be evolved than when they can't. Further, for both conditions, there were fewer good learners but more moderate learners in the case where instincts could be evolved. These data seem to suggest that the possibility of evolving instincts had neither an inhibiting nor an enhancing effect on the evolution of learning but, rather, a moderating effect, with fewer good learners and fewer non-substantial learners but more moderate learners evolved when instincts were possible. It is interesting that the data from both the nurturing and self-care niches seem to reflect this same effect. However, the size of the differences found were quite small and far from statistical significance, so we will base no conclusions on this data and instead leave further consideration of this question as future research.

Looking at the differences in performance continuum results, we see that within each category of learner it appears that the possibility of evolving instincts improves the performance of the individuals involved. However, these differences are also quite small and are not statistically significant, so we will likewise base no conclusions on this data.

7 Conclusions

This work extends and builds on prior work in the artificial evolution of machine learning, particularly within artificial neural networks (Soltoggio et al., 2017). In particular, it investigates the hypothesis that nurturing promotes the evolution of learning, an idea that has been pursued in our lab for several years (Woehrer et al., 2012) and is beginning to find a larger audience (Soltoggio et al., 2017).

We considered the evolution of learning in the presence of nurturing by task simplification. The overall impact of nurturing by task simplification is similar to that of reward shaping (Laud, 2004; Norouzzadeh, 2010) in the sense that learning is observed more often if the task is simpler.

The results suggest that niche construction changes the dynamics of the evolutionary process as seen by the nurturing niche outperforming the self-care niche in terms of the evolution of learning. The statistical tests indicate strongly that nurturing promotes the evolution of learning in changing environments.

This is the first study to demonstrate that nurturing promotes the evolution of the components of learning in changing environments even in the presence of instincts. In contrast to this work, Eskridge & Hougen (2012) used an abstract environment with no evolution of learning rule parameters. Moreover, they considered nurturing as either safe exploration or social learning rather than as task simplification. However, the results in this study conform to their claim that "nurturing promotes the evolution of learning in uncertain environments in which learning would otherwise not be a viable strategy at statistically significant levels."

Relatedly, Hoke (2017) also considers whether nurturing promotes the evolution of learning. However, that work considers nurturing as safe exploration (as with Eskridge & Hougen, 2012) and the evolution of supervised learning rather than reinforcement learning. Nonetheless, that work also concluded that nurturing promotes the evolution of learning, adding to the body of work supporting this overarching hypothesis.

Hoke also considers whether the evolvability of instincts influences the evolution of learning. In that study, it was found that allowing instincts to be evolved could inhibit the evolution of learning. Here, however, we find no evidence for that conclusion, though no evidence against that conclusion either. Instead, our data seem to suggest that the possibility of evolving instincts moderates the type of learning evolved while boosting overall agent performance. It must be stressed, however, that our results here on these points is not statistically significant and therefore not conclusive.

This work contributes to understanding the virtuous cycle (see Figure 1) by connecting the evolution of nurturing (Leonce et al., 2012) to the second half of the cycle, learning to be a better nurturer. In biology, nurturing and learning are studied separately; however, machine learning provides us a platform to integrate the two with the objective to develop more robust algorithms that can solve arbitrarily complex tasks with more flexibility. In order to develop better machine learning algorithms, this work points out nurturing as an important part of the solution space where robots learn to perform complex tasks.

8 Future Work

This work is a contribution to the new area of nurturing robotics. The future seems promising based on results shown for the evolution of nurturing (Leonce et al., 2012), results shown that support the broad hypothesis that nurturing promotes the evolution of learning (Eskridge & Hougen, 2012; Shah & Hougen, 2017a; Hoke, 2017), and this work that demonstrates that nurturing promotes the evolution of the components of reinforcement learning in changing environments regardless of instincts. This section discusses the next major steps in this research agenda.

8.1 Learning to be a Better Nurturer

The next obvious step in this research is to close the loop in the proposed virtuous cycle (Figure 1) by demonstrating that learning enables greater nurturing. One way to do this would be to apply various successful learning algorithms evolved in the present experiments to a parent to find out if it can learn to be a better nurturer for its offspring. This would mean, for example, that an arena could be designed in which there are several light switches on one end that activate a single light source on the other end. The reward value of the light would be determined by which switch is turned on by the parent/nurturer. The parent's job would be to choose the right switch to turn on in order to provide maximum reward to its offspring. That would also mean that a communication mechanism between the child and the parent must exist and, preferably, be evolved. Based on the feedback from the child, the parent should improve on its behavior and make intelligent decisions over its lifetime. A reward switch between lifetimes and during each lifetime will be essential again to make this arena a non-stationary environment to encourage learning. This experiment should

also allow us to validate how general, robust, and scalable the evolved learning algorithms are, as well as help us understand whether the evolution of learning in turn enables an individual to be a better nurturer, thus completing the virtuous cycle.

8.2 Lamarckian Inheritance

Another area that can be explored is Lamarckian inheritance. *Lamarckian inheritance* is the inheritance of acquired characteristics (Kronfeldner, 2006). It will be interesting to find out if nurturing promotes the evolution of learning using Lamarckian inheritance as well. For example, an individual neural controller could be allowed to retain its learned synaptic weight means and pass them on to its offspring, which could be considered a form of Lamarckian inheritance.

Using this setup, another interesting question that can be investigated is how Lamarckian inheritance interacts with environmental change and the evolution of learning. Consider, for example, two contrasting environments. In the first, the environment changes toward the end of an individual's lifetime but not between generations. In the second, the environment likewise changes toward the end of an individual's lifetime, but then changes back to its original state between generations. In the first environment, Lamarckian inheritance might promote the evolution of learning because an individual's learning would benefit both itself (since its learning helps it to adapt to its environment after the change) and its child (since what it learns is passed on to its child, which begins its own lifetime in a similar environment). However, in the second environment, Lamarckian inheritance might interfere with the evolution of learning because, while an individual's learning would benefit the individual itself (again, it would help to adapt the individual to its environment after the change), it might hamper its child, which inherits behavior that is sub-optimal for most of its lifetime. Here, an individual might be better off acting instinctively throughout its lifetime so long as those instincts served it well for most of its lifetime and likewise helped its offspring for most of its lifetime.

This could also be seen as a model of an alternative form of nurturing, as parental knowledge sharing with offspring (for example, through instruction or demonstration) is not so different from Lamarckian inheritance of learned knowledge except, of course, that cultural knowledge isn't passed through genes.

8.3 Risk Analysis

By introducing reward variability to the environment, risk analysis can be performed (as in Niv et al., 2002). Consider a single switch, multi-light environment as used in the present work. In this environment, risk aversion can be analyzed using an experimental setup such as the following:

1. Very risky resource: A high-rewarding light source with a 1.0 reward 10% of the time and 0.0 the other 90% of the time.
2. Risky resource: A medium-rewarding light source with a 0.5 reward 20% of the time and 0.0 the other 80% of the time.
3. Non-risky resource: A low-rewarding light source with a 0.1 reward all the time consistently.

This will give three different resource options, all with the same mean reward (0.1) but with different levels of variance—a standard form of the more general concept of risk in economics (Rothschild & Stiglitz, 1970). Because risk aversion is a side effect of reinforcement learning (Niv et al., 2002), individuals should learn to visit the non-risky resource most often. One question that could be asked is whether reinforcement learning would be as readily evolved in an environment characterized by differences in reward risk (in this case, variability around a common mean reward) rather than by differences in mean reward. Another question that could be asked is whether nurturing promotes the evolution of learning in environments characterized by differences in reward risk. An element that can be added to each of these questions is whether reinforcement learning is evolved as readily when riskier options have higher mean rewards than less risky options.

8.4 Instincts

While both this work and that of Hoke (2017) demonstrate that nurturing promotes the evolution of learning regardless of whether instincts can also be evolved, the overall effect

of allowing instincts to be evolved is unclear. In the work of Hoke, allowing instincts to evolve was found to substantially inhibit the evolution of learning while this work shows no such effect and instead hints at a moderating effect on the type of learning evolved and an enhanced effect on the quality of overall behavior evolved. Clearing up this picture warrants further research.

Acknowledgements

The authors gratefully acknowledge the substantial supercomputing resources, including outstanding personal assistance, provided by the OU Supercomputing Center for Education & Research (OSCER), without which this research would not have been possible.

Author Biographies

To typeset an "Author Biographies" section.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. For a complete list of these sub-hypotheses, see Shah (2015), Section 4.1.2.
2. There is, of course, the question of how evolution arrives at that optimal policy and the answer is that the evolutionary process itself must contain an exploratory component (e.g., mutation). This means that so long as the evolutionary process continues, an individual is not guaranteed to have optimal instincts even if its immediate ancestor(s) did. In this case, learning might still be valuable even if the environment external to the evolved lineage is unchanging.
3. Contrast this figure with the one from Ellefsen (2013) that does not explicitly represent the possibility of overlap between instincts and learning but does consider sensitive periods for learning within a lifetime versus lifelong learning. Also, note that a spectrum of change is only one possible way to conceive of change in the environment and its possible effect on the evolution of learning. For example, one might consider fixity and reliability independently, where *fixity* is the tendency of an environment to resist long-term changes and *reliability* is the tendency of an environment to resist short-term fluctuations (Dunlap & Stephens, 2009; Eskridge & Hougen, 2012; Ellefsen, 2014). Here, reliability would be related to the economic definition of risk (Rothschild & Stiglitz, 1970).
4. While the left and the right light positions are the same distance from the robot's starting position in the arena, the robot's starting orientation means that the minimum time to the right light position is longer than the minimum time to the left light position. (See Figure 3 for robot orientation.)
5. Note that linear regression lines and slopes in the graphs are presented solely to highlight trends in the data. They are not used to determine to which category an example repetition belongs.
6. Note that the possibility of a multiple comparisons problem exists any time multiple statistical hypothesis tests are naïvely applied. For a discussion of the multiple comparisons problem as it applies to the other statistical hypothesis tests found herein, see Shah (2015), Section 4.1.3.

References

Amarjyoti, S. (2017). Deep reinforcement learning for robotic manipulation: The state of the art. *arXiv preprint arXiv:1701.08878*.

Anderson, R. W. (1995). Learning and evolution: A quantitative genetics approach. *Journal of Theoretical Biology*, *175(1)*, 89–101.

Aoki, K. & Feldman, M. W. (2014). Evolution of learning strategies in temporally and spatially variable environments: A review of theory. *Theoretical Population Biology*, *91*, 3–19.

Argall, B. D., Chernova, S., Veloso, M. & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57(5)*, 469–483.

Baxter, J. (1993). The evolution of learning algorithms for artificial neural networks. *Complex Systems* (pp. 313–326).

Behere, S. (2015). Architecture support for automobile autonomy: A state of the art survey. Technical Report

KTH TRITA - MMK 2015:08, Kungliga Tekniska Högskolan (KTH).

Bekey, G., Ambrose, R., Kumar, V., Lavery, D., Sanderson, A., Wilcox, B., Yuh, J. & Zheng, Y. (2008). *Robotics: State of the Art and Future Challenges*. World Scientific.

Bekey, G. A. (1996). Editorial: On autonomy and intelligence. *Autonomous Robots*, *3(5)*, 1 page.

Bengio, S., Bengio, Y., Cloutier, J. & Gecsei, J. (1992). On the optimization of a synaptic learning rule. In *Optimality in Artificial and Biological Neural Networks* (pp. 6–8). Univ. of Texas.

Bolles, R. C. & Beecher, M. D. (Eds.). (1988). *Evolution and Learning*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Bongard, J. (2011). The 'what', 'how' and the 'why' of evolutionary robotics. In *New Horizons in Evolutionary Robotics*, Studies in Computational Intelligence (pp. 29–35). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-18272-3_2.

Bongard, J. C. (2013). Evolutionary robotics. *Communications of the ACM*, *56(8)*, 74–3.

Botvinick, M. M., Niv, Y. & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113(3)*, 262–280.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34(2)*, 181–186.

Campbell, N. A. & Reece, J. B. (2008). *Biology* (8th Ed.). Benjamin Cummings.

Chalmers, D. J. (1990). The evolution of learning: An experiment in genetic connectionism. In *Proceedings of the 1990 Connectionist Models Summer School* (pp. 81–90). Morgan Kaufmann.

Char, K. G. (1997). Evolution of structure and learning: A GP approach. In J. Mira, R. Moreno-Díaz & J. Cabestany (Eds.), *Biological and Artificial Computation: From Neuroscience to Technology*, Volume 1240 of *Lecture Notes in Computer Science* (pp. 510–517). Springer Berlin Heidelberg.

Chen, X., Chen, Y. Q. & Chase, J. G. (Eds.). (2009). *Mobile Robots—State of the Art in Land, Sea, Air, and Collaborative Missions*. InTech.

Chernova, S. & Thomaz, A. L. (2014). Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *8(3)*, 1–121.

Commuri, S., Fierro, R., Hougen, D. & Muthuraman, R. (2004). System intelligence requires distributed learning. In *Proceedings of the 2004 IEEE International Symposium on Intelligent Control, 2004.* (pp. 67–72). IEEE.

Dasdan, A. & Oflazer, K. (1993). Genetic synthesis of unsupervised learning algorithms. Technical Report BU-CEIS-9305, Department of Computer Engineering and Information Science, Bilkent University.

Daw, N. D. & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, *113(3)*, 259–261.

Dayan, P. & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18(2)*, 185–196.

Di Paolo, E. A. (2003). Evolving spike-timing-dependent plasticity for single-trial learning in robots. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *361(1811)*, 2299–2319.

Doelling, K., Shin, J. & Popa, D. O. (2014). Service robotics for the home: A state of the art review. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '14 (pp. 35:1–35:8). New York, NY, USA: ACM.

Doll, B. B., Jacobs, W. J., Sanfey, A. G. & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. Computational Cognitive Neuroscience II.

Doncieux, S., Bredeche, N., Mouret, J.-B. & Eiben, A. E. G. (2015). Evolutionary robotics: What, why, and where to. *Frontiers in Robotics and AI*, *2*, 1–18.

Dridi, S. & Lehmann, L. (2014). On learning dynamics underlying the evolution of learning rules. *Theoretical Population Biology*, *91*, 20–36.

Dudek, G. & Jenkin, M. (2010). *Computational Principles of Mobile Robotics* (Second Edition Ed.). Cambridge University Press.

Dukas, R. (1998). Evolutionary ecology of learning. In R. Dukas (Ed.), *Cognitive Ecology: The Evolutionary Ecology of Information Processing and Decision Making* (pp. 129–174). Chicago: University of Chicago Press.

Dukas, R. (2013). Effects of learning on evolution: Robustness, innovation and speciation. *Animal Behaviour*, *85(5)*, 1023–1030. Including Special Section: Behavioural Plasticity and Evolution.

Dunlap, A. S. & Stephens, D. W. (2009). Components of change in the evolution of learning and unlearned preference. *Proceedings of the Royal Society B: Biological Sciences*, *276(1670)*, 3201–3208.

Eiben, A. E. (2014). Grand challenges for evolutionary robotics. *Frontiers in Robotics and AI*, *1*, 1–2.

Eiben, A. E. & Smith, J. E. (2015). Evolutionary robotics. In *Introduction to Evolutionary Computing*, Natural Computing Series (pp. 245–258). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-662-44874-8_17.

Ellefsen, K. (2013). Balancing the costs and benefits of learning ability. In *12th European Conference on Artificial Life (ECAL 2013)* (pp. 292–299). MIT Press.

Ellefsen, K. (2014). The evolution of learning under environmental variability. In *Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems (ALife 14)* (pp. 649–656). The MIT Press.

Eskridge, B. E. & Hougen, D. F. (2012). Nurturing promotes the evolution of learning in uncertain environments. In *2012 IEEE International Conference on Development and Learning (ICDL) and Epigenetic Robotics* (p. 6 pages). IEEE.

Federici, D. (2005). Evolving developing spiking neural networks. In *2005 IEEE Congress on Evolutionary Computation*, Volume 1 (pp. 543–550). IEEE.

Feldman, M. W. & Laland, K. N. (1996). Gene-culture coevolutionary theory. *Trends in Ecology & Evolution*, *11(11)*, 453–457.

Floreano, D., Dürr, P. & Mattiussi, C. (2008). Neuroevolution: From architectures to learning. *Evolutionary Intelligence*, *1(1)*, 47–62.

Floreano, D. & Urzelai, J. (2000). Evolutionary robots with on-line self-organization and behavioral fitness. *Neural Networks*, *13(4-5)*, 431–443.

Fonseca, I. M. d. & Pontuschka, M. N. (2015). The state-of-the-art in space robotics. *Journal of Physics: Conference Series*, *641(1)*, 8 pages.

Fontanari, J. F. & Meir, R. (1991). Evolving a learning algorithm for the binary perceptron. *Network: Computation in Neural Systems*, *2(4)*, 353–359.

Ghanbari, A., Vaghei, Y., Noorani, S. & Reza, S. M. (2014). Reinforcement learning in neural networks: A survey. *International Journal of Advanced Biological and Biomedical Research*, *2(5)*, 1398–1416.

Haasdijk, E., Bredeche, N., Nolfi, S. & Eiben, A. E. (2014). Evolutionary robotics. *Evolutionary Intelligence*, *7(2)*, 69–70.

Harley, C. B. & Smith, J. M. (1983). Learning—evolutionary approach. *Trends in Neurosciences*, *6*, 204–208.

van Hasselt, H. (2012). Reinforcement learning in continuous state and action spaces. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art* (pp. 207–251). Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-27645-3_7.

Heller, D. (2004). An evolutionary approach to learning in a changing environment. *Journal of Economic Theory*, *114(1)*, 31–55.

Hernandez, C., Bharatheesha, M., Ko, W., Gaiser, H., Tan, J., van Deurzen, K., ... & Wisse, M. (2016). Team Delft's robot winner of the Amazon Picking Challenge 2016. *arXiv:1610.05514 [cs]*. arXiv: 1610.05514.

Hexmoor, H. (2013). *Essential Principles for Autonomous Robotics*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.

Hoke, B. (2017). Nurturing as safe exploration promotes the evolution of generalized supervised learning. Master's thesis, University of Oklahoma.

Huntsberger, T., Keegan, M. & Brizzolara, R. (2010). Editorial: Special issue on state of the art in maritime autonomous surface and underwater vehicles, part 1. *Journal of Field Robotics*, *27(6)*, 699–701.

Huntsberger, T., Keegan, M. & Brizzolara, R. (2011). Editorial: Special issue on state of the art in maritime autonomous surface and underwater vehicles, part 2. *Journal of Field Robotics*, *28(1)*, 1–2.

Husbands, P. (2017). Evolutionary robotics. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning*

and Data Mining (pp. 469–480). Springer US. DOI: 10.1007/978-1-4899-7687-1_94.

Ingrand, F. & Ghallab, M. (2017). Deliberation for autonomous robots: A survey. *Artificial Intelligence*, *247*, 10–44.

Johnston, T. D. (1982). Selective costs and benefits in the evolution of learning. In C. B. Jay S. Rosenblatt, Robert A. Hinde & M.-C. Busnel (Eds.), *Advances in the Study of Behavior*, Volume 12 of *Advances in the Study of Behavior* (pp. 65–106). Academic Press.

Kapach, K., Barnea, E., Mairon, R., Edan, Y. & Ben-Shahar, O. (2012). Computer vision for fruit harvesting robots—State of the art and challenges ahead. *International Journal of Computational Vision and Robotics*, *3(1-2)*, 4–34.

Kendal, J., Giraldeau, L.-A. & Laland, K. (2009). The evolution of social learning rules: Payoff-biased and frequency-dependent biased transmission. *Journal of Theoretical Biology*, *260(2)*, 210–219.

Kerr, B. & Feldman, M. W. (2003). Carving the cognitive niche: Optimal learning strategies in homogeneous and heterogeneous environments. *Journal of Theoretical Biology*, *220(2)*, 169–188.

Kirchkamp, O. (1999). Simultaneous evolution of learning rules and strategies. *Journal of Economic Behavior & Organization*, *40(3)*, 295–312.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4(3–4)*, 415–452.

Kober, J., Bagnell, J. A. & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32(11)*, 1238–1274.

Kober, J. & Peters, J. (2012). Reinforcement learning in robotics: A survey. In *Reinforcement Learning: State-of-the-Art* (pp. 579–610). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kober, J. & Peters, J. (2014a). *Learning Motor Skills*, Volume 97 of *Springer Tracts in Advanced Robotics*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-03194-1.

Kober, J. & Peters, J. (2014b). Reinforcement learning in robotics: A survey. In *Learning Motor Skills*, Volume 97 (pp. 9–67). Cham: Springer International Publishing. DOI: 10.1007/978-3-319-03194-1_2.

Kormushev, P., Calinon, S. & Caldwell, D. G. (2013). Reinforcement learning in robotics: Applications and real-world challenges. *Robotics*, *2(3)*, 122–148.

Krakauer, D. C. & Rodríguez-Gironés, M. A. (1995). Searching and learning in a random environment. *Journal of Theoretical Biology*, *177(4)*, 417–429.

Kronfeldner, M. E. (2006). Is cultural evolution Lamarckian? *Biology & Philosophy*, *22(4)*, 493–512.

Kronreif, G. (2016). Medical robotics—State-of-the-art and future trends. In *2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES)* (pp. 17–18). IEEE.

Laud, A. D. (2004). *Theory and Application of Reward Shaping in Reinforcement Learning*. PhD thesis, University of Illinois at Urbana-Champaign.

Lehman, J. & Miikkulainen, R. (2014). Overcoming deception in evolution of cognitive behaviors. In *Genetic and Evolutionary Computation Conference (GECCO)* (pp. 185–192). ACM Press.

Leonce, A., Hoke, B. & Hougen, D. F. (2012). Evolution of robot-to-robot nurturing and nurturability. In *2012 IEEE International Conference on Development and Learning (ICDL) and Epigenetic Robotics* (p. 7 pages). IEEE.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274v3* (p. 66 pages).

Liu, Y. & Nejat, G. (2013). Robotic urban search and rescue: A survey from the control perspective. *Journal of Intelligent & Robotic Systems*, *72(2)*, 147–165.

Lotem, A. & Biran-Yoeli, I. (2014). Evolution of learning and levels of selection: A lesson from avian parent–offspring communication. *Theoretical Population Biology*, *91*, 58–74.

Magnenat, S., Waibel, M. & Beyeler, A. (2007). Enki, a fast 2D robot simulator. http://home.gna.org/enki/.

Marques, L., De Almeida, A. T., Armada, M., Fernández, R., Montes, H., González, P. & Baudoin, Y. (2012). State of the art review on mobile robots and manipulators for humanitarian demining. *IARP WS on Humanitarian Demining* (pp. 2–8).

Maurer, J., Steinbauer, G., Lepej, P. & Uran, S. (2014). TEDUSAR white book—State of the art in search and rescue robots. *Graz University of Technology, University of Maribor, Tech. Rep.*

Mirbagheri, A., Baniasad, M. A., Farahmand, F., Behzadipour, S. & Ahmadian, A. (2013). Medical robotics: State-of-the-art applications and research challenges. *International Journal of Healthcare Information Systems and Informatics*, *8(2)*, 1–14.

Mondada, F., Bonani, M., Raemy, X., Pugh, J., Cianci, C., Klaptocz, A., Magnenat, S., Zufferey, J.-C., Floreano, D. & Martinoli, A. (2009). The e-puck, a robot designed for education in engineering. In Gonçalves, P. J. S., Torres, P. J. D. & Alves, C. M. O. (Eds.), *Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions*, Volume 1 (pp. 59–65). Portugal: IPCB: Instituto Politécnico de Castelo Branco.

Moore, B. R. (2004). The evolution of learning. *Biological Revue*, *79*, 301–335.

Murphy, R. R. (2015). Meta-analysis of autonomy at the DARPA robotics challenge trials. *Journal of Field Robotics*, *32(2)*, 189–191.

Nakahashi, W. (2007). The evolution of conformist transmission in social learning when the environment changes periodically. *Theoretical Population Biology*, *72(1)*, 52–66.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53(3)*, 139–154. Special Issue: Dynamic Decision Making.

Niv, Y., Joel, D., Meilijson, I. & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behaviour*, *10(1)*, 5–24.

Nolfi, S., Bongard, J., Husbands, P. & Floreano, D. (2016). Evolutionary robotics. In *Springer Handbook of Robotics* (pp. 2035–2068). Springer, Cham. DOI: 10.1007/978-3-319-32552-1_76.

Nolfi, S. & Parisi, D. (1996). Learning to adapt to changing environments in evolving neural networks. *Adaptive Behaviour*, *5(1)*, 75–98.

Norouzzadeh, S. (2010). *Shaping Methods to Accelerate Reinforcement Learning: From Easy to Challenging Tasks*. PhD thesis, Delft University of Technology.

Orchard, J. & Wang, L. (2016). The evolution of a generalized neural learning rule. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 4688–4694). IEEE.

Ott, L. & Ramos, F. (2013). Unsupervised online learning for long-term autonomy. *The International Journal of Robotics Research*, *32(14)*, 1724–1741.

Pagliuca, P. & Nolfi, S. (2015). Integrating learning by experience and demonstration in autonomous robots. *Adaptive Behavior* (pp. 300–314).

Peters, J., Kober, J., Mülling, K., Krämer, O. & Neumann, G. (2013). Towards robot skill learning: From simple skills to table tennis. In Blockeel, H., Kersting, K., Nijssen, S. & Železný, F. (Eds.), *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III* (pp. 627–631). Berlin, Heidelberg: Springer Berlin Heidelberg.

Peters, J., Morimoto, J., Tedrake, R. & Roy, N. (2009). Robot learning [TC spotlight]. *IEEE Robotics Automation Magazine*, *16(3)*, 19 –20.

Plotkin, H. & Odling-Smee, F. (1979). Learning, change, and evolution: An enquiry into the teleonomy of learning. In J. S. Rosenblatt, R. A. Hinde, C. Beer & M.-C. Busnel (Eds.), *Advances in the Study of Behavior*, Volume 10 (pp. 1–41). Academic Press.

Polydoros, A. S. & Nalpantidis, L. (2017). Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, *86(2)*, 153–173.

Risi, S., Hughes, C. E. & Stanley, K. O. (2010). Evolving plastic neural networks with novelty search. *Adaptive Behavior*, *18(6)*, 470–491.

Risi, S. & Stanley, K. O. (2010). Indirectly encoding neural plasticity as a pattern of local rules. In *From Animals to Animats 11* (pp. 533–543). Springer.

Risi, S. & Stanley, K. O. (2012). A unified approach to evolving plasticity and neural geometry. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Risi, S., Vanderbleek, S. D., Hughes, C. E. & Stanley, K. O. (2009). How novelty search escapes the deceptive trap of learning to learn. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, GECCO '09 (pp. 153–160). New York, NY, USA: ACM.

Rothschild, M. & Stiglitz, J. E. (1970). Increasing risk I: A definition. *Journal of Economic Theory* (pp. 225–243).

Sales, J., Santos, L., Sanz, P. J., Dias, J. & García, J. C. (2014). Increasing the autonomy levels for underwater intervention missions by using learning and probabilistic techniques. In *ROBOT2013: First Iberian Robotics Conference*, Advances in Intelligent Systems and Computing (pp. 17–32). Springer, Cham. DOI: 10.1007/978-3-319-03413-3_2.

Shah, A. (2012). Psychological and neuroscientific connections with reinforcement learning. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art* (pp. 507–537). Berlin, Heidelberg: Springer Berlin Heidelberg.

Shah, S. N. H. (2015). *Nurturing Promotes the Evolution of Learning in Changing Environments*. PhD thesis, University of Oklahoma, Norman, OK, USA.

Shah, S. N. H. & Hougen, D. F. (2017a). Nurturing promotes the evolution of reinforcement learning in changing environments. In *2017 IEEE Symposium on Computational Intelligence in Control and Automation (CICA)*. IEEE. To appear.

Shah, S. N. H. & Hougen, D. F. (2017b). The stochastic synapse reinforcement learning algorithm. In *2017 IEEE Symposium on Computational Intelligence in Control and Automation (CICA)*. IEEE. To appear.

Shteingart, H. & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, *25*, 93–98.

Siegwart, R., Nourbakhsh, I. R. & Scaramuzza, D. (2011). *Introduction to Autonomous Mobile Robots* (Second Ed.). Intelligent Robotics and Autonomous Agents. Cambridge, MA, USA: MIT Press.

Silva, F., Correia, L. & Christensen, A. L. (2016). Evolutionary robotics. *Scholarpedia*, *11(7)*, 33333.

Snell-Rood, E. C. (2013). An overview of the evolutionary causes and consequences of behavioural plasticity. *Animal Behaviour*, *85(5)*, 1004–1011. Including Special Section: Behavioural Plasticity and Evolution.

Soltoggio, A. (2008). Neural plasticity and minimal topologies for reward-based learning. In *Eighth International Conference on Hybrid Intelligent Systems, 2008 (HIS'08)* (pp. 637–642). IEEE.

Soltoggio, A., Bullinaria, J. A., Mattiussi, C., Dürr, P. & Floreano, D. (2008). Evolutionary advantages of neuromodulated plasticity in dynamic, reward-based scenarios. In *Proceedings of the 11th International Conference on Artificial Life (Alife XI)* (pp. 569–576). MIT Press.

Soltoggio, A., Dürr, P., Mattiussi, C. & Floreano, D. (2007). Evolving neuromodulatory topologies for reinforcement learning-like problems. In *IEEE Congress on Evolutionary Computation* (pp. 2471–2478). IEEE.

Soltoggio, A. & Jones, B. (2009). Novelty of behaviour as a basis for the neuro-evolution of operant reward learning. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation* (pp. 169–176). ACM.

Soltoggio, A., Stanley, K. O. & Risi, S. (2017). Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks. *arXiv preprint arXiv:1703.10371*.

Stephens, D. W. (1991). Change, regularity, and value in the evolution of animal learning. *Behavioral Ecology*, *2(1)*, 77–89.

Stone, P. & Veloso, M. M. (2000). Layered learning. In *European Conference on Machine Learning*, Volume 1810 of *Lecture Notes in Computer Science* (pp. 369–381). Springer.

Sukhatme, G. (Ed.). (2009). *The Path to Autonomous Robots*. Boston, MA: Springer US. DOI: 10.1007/978-0-387-85774-9.

Taylor, A. & Riek, L. D. (2016). Robot perception of human groups in the real world: State of the art. In *AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction Technical Report FS-16-01*, Volume 4 (p. 2017). AAAI.

Theodoridis, T. & Hu, H. (2012). Toward intelligent security robots: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42(6)*, 1219–1230.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., ... & Mahoney, P. (2007). Stanley: The robot that won the DARPA Grand Challenge. In M. Buehler, K. Iagnemma & S. Singh (Eds.), *The 2005 DARPA Grand Challenge*, Volume 36 of *Springer Tracts in Advanced Robotics* (pp. 1–43). Springer.

Vargas, P. A., Di Paolo, E. A., Harvey, I. & Husbands, P. (Eds.). (2014). *The Horizons of Evolutionary Robotics*. Intelligent Robots and Autonomous Agents. Cambridge, Massachusetts: The MIT Press.

Walter, W. G. (1951). A machine that learns. *Scientific American*, *185(2)*, 60–63.

Walter, W. G. (1953). *The Living Brain*. Norton.

Wiering, M. & van Otterlo, M. (Eds.). (2012). *Reinforcement Learning*, Volume 12 of *Adaptation, Learning, and Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-27645-3.

Woehrer, M., Hougen, D. F., Schlupp, I. & Eskridge, B. E. (2012). Robot-to-robot nurturing: A call to the research community. In *Joint IEEE International Conference on Development and Learning (ICDL) and Epigenetic Robotics* (p. 2 pages). IEEE.

Yanco, H. A., Norton, A., Ober, W., Shane, D., Skinner, A. & Vice, J. (2015). Analysis of human-robot interaction at the DARPA Robotics Challenge trials. *Journal of Field Robotics*, *32(3)*, 420–444.

Zuzánek, P., Zimmermann, K. & Hlaváč, V. (2014). Accepted autonomy for search and rescue robotics. In Hodicky, J. (Ed.), *Modelling and Simulation for Autonomous Systems: First International Workshop, MESAS 2014, Rome, Italy, May 5-6, 2014, Revised Selected Papers* (pp. 231–240). Cham: Springer International Publishing.