Coevolutionary Learning of Neuromodulated Controllers for Multi-Stage and Gamified Tasks

Chloe M. Barnes*, Anikó Ekárt*, Kai Olav Ellefsen[†], Kyrre Glette[†‡], Peter R. Lewis* and Jim Tørresen[†‡]
*Department of Computer Science, Aston University, Birmingham, UK
{barnecm1, a.ekart, p.lewis}@aston.ac.uk
[†]Department of Informatics, University of Oslo, Oslo, Norway
{kaiolae, kyrrehg, jimtoer}@ifi.uio.no
[‡]RITMO, University of Oslo, Oslo, Norway

***Abstract*—Neural networks have been widely used in agent learning architectures; however, learning multiple context-dependent tasks simultaneously or sequentially is problematic when using them. Behavioural plasticity enables humans and animals alike to respond to changes in context and environmental stimuli, without degrading learnt knowledge; this can be achieved by regulating behaviour with *neuromodulation* – a biological process found in the brain. We demonstrate that modulating activity-propagating signals when evolving neural networks enables agents to learn context-dependent and multi-stage tasks more easily. Further, we show that this benefit is preserved when agents occupy an environment shared with other neuromodulated agents. Additionally we show that neuromodulation helps agents that have evolved alone to adapt to changes in environmental stimuli when they continue to evolve in a shared environment.**

I. Introduction

Natural and artificial environments are often complex, unpredictable and dynamic, making learning and surviving a challenge for both animals and artificial agents alike [1], [2]. In order to survive in these challenging conditions, many organisms such as nematodes [3], fish [4] and the African striped mouse *Rhabdomys pumilio* [5] show phenotypic, activational and behavioural plasticity; this ability to express different behaviours – and reverse them depending on varying environmental stimuli – allows rapid adaptation to novel situations [5], [6].

As with animals in the real world, controllers tasked with learning in dynamic and unpredictable artificial environments – artificial neural networks (ANNs) in particular – face similar challenges. Neuroevolution is the process of evolving ANNs with an evolutionary algorithm in accordance to a fitness function [7]; a population of individuals is evolved with mutations and/or crossover over many generations [8]. Many applications of neuroevolution focus on evolving weights of ANNs [8]–[11], however more complex approaches that evolve both the weights and topologies of ANNs exist [12], [13]. This process of evolving ANNs by adjusting connection weights over time to encode new information can result in a degradation of performance and catastrophic forgetting when learning new tasks or experiencing novel environmental contexts [8], [14]–[16]; learnt knowledge must be changed – and is often lost – in order to learn new things and express new behaviours [8]. Learning complex, sequential or multi-stage tasks is also made difficult as complete information about the environment – including the available actions, their cues and their consequences – is not usually accessible [1], [17]; this is also evident when environments are shared, as the actions of individuals change the context of the environment for others [11].

In nature, the immediate and reversible behavioural changes as a result of behavioural plasticity that facilitate adaptations to novel contexts can be achieved with *neuromodulation* – a biological process whereby chemical signals are gated in the brain depending on environmental stimuli and situations [18]. Consequently, neuromodulation has thus been used to aid neural controllers with learning new or sequential tasks, and learning in dynamic environments [8], [19], [20].

We explore how activity-gating neuromodulation may help neural controllers to overcome the challenges associated with learning multi-stage and gamified tasks, without a priori knowledge of the task or environment. ANNs are just one example of an agent controller in which behaviour can be learnt; we use ANNs in this context in line with previous River Crossing testbeds [9]–[11], to explore how ANNs make decisions in social environments. Here, a multi-stage task is defined as one that an agent must learn, and pass, through multiple states, and perform different behaviours in different contexts in order to achieve their goal; this definition is inspired by [17]. Further, we investigate how this regulation of behaviour may help these agents to learn in multi-agent environments, without the capacity to learn of the existence of others; the act of introducing other agents to the environment changes the context of the task, which becomes an implicit social dilemma. Neuromodulation has been used to explore social dynamics in multi-agent systems [2], [21], however our work extends the notion of [11], where cooperation and exploitation may be emergent behaviours, but cannot be *intentional*. We hypothesise that reversible and immediate behavioural plasticity as a result of neuromodulation will enable agents to better learn the state-space of their environment and therefore the task at hand. We demonstrate this using the River Crossing Dilemma (RCD) testbed introduced by [11], as well as a new adaptation of the environment called the *Protected* River Crossing Dilemma (PRCD), which we introduce to explore and contrast how agents learn to solve *single*-stage tasks under these same conditions.

II. Background

A. Behavioural Plasticity and Neuromodulation

One way to design adaptive systems is to look at the theory of behavioural plasticity. Behavioural plasticity can be seen as the ability to change or adapt behaviour based on changes in environmental stimuli [22]; this is important for navigating uncertain, novel or dynamic environments and can be classed into two different types: developmental and activational [6]. Developmental behavioural plasticity can be seen as learning from experience and external stimuli. Activational behavioural plasticity on the other hand enables immediate behavioural changes; individuals can respond to new or dynamic environments during their lifetime by changing their phenotype. Activational plasticity is also termed 'innate' [22] or 'contextual' [23] plasticity.

Neuromodulation is a biological process found in animal brains [24], whereby chemical signals modify, gate or regulate synaptic plasticity based on the modulatory signal combined with the pre- and post-synaptic activities, and environmental stimuli [8], [18], [25]. In neuroscience, synaptic plasticity is the modification of synapses between neurons through strengthening or weakening them [26]. In ANNs, synaptic plasticity is achieved by modulating neural network weights; short-term modifications result in immediate phenotypic changes, and long-term changes result in learning and adaptation based on experience. Developmental plasticity is achieved by regulating learning in the long-term, where modulatory signals alter synaptic strengths; activational plasticity is achieved by regulating behaviour or synaptic activity in the short-term with the modulatory signal, without affecting learning and without long-lasting changes to synaptic strengths.

B. Achieving Developmental Plasticity with Neuromodulation

Similarly to ANNs being inspired by the connectionist architectures found in brains, neuromoduation has been widely applied to artificial models to regulate synaptic plasticity and the learning rate of neural connections. Neural networks have been evolved with modulatory neurons to regulate learning and mitigate the catastrophic forgetting associated with performing tasks in uncertain environments [25]; this method has been found to improve learning in T-maze problems. Other studies have found that promoting the evolution of modular neural networks by introducing a cost for neural connections can mitigate catastrophic forgetting and improve learning; here, learning is regulated with neuromodulation [8]. Neuromodulation has also been used to develop conflict learning in neural networks [27], and associative learning in real robots [28]; these two approaches employ neuromodulation, but do not use neuroevolution as a learning mechanism.

The approaches outlined in this section modulate learning and therefore developmental plasticity by regulating the local learning rate of neurons in the network; they do not however demonstrate how behaviour can be regulated in a short-term, reversible way *without* affecting learning, in order to facilitate *immediate* behavioural changes to changing environmental stimuli. Further, these approaches only use neuromodulation in neural networks or robots that exist in isolation; we however explore how immediate behavioural plasticity can be achieved with neuromodulation in agents without regulating learning, in single- and also multi-agent environments.

C. Achieving Activational Plasticity with Neuromodulation

Neurobiological mechanisms have been explored using a computational framework based on neuromodulatory systems such as the dopaminergenic and serotonergic systems, by regulating synaptic activity [29]. Whilst this is proposed to aid autonomous agents in exploratory and exploitative decision making, activational plasticity is not applied as a tool to improve neuroevolution, but rather to model and explore biological systems computationally. The effects of modulating neuroreceptors and synaptic plasticity have been studied with spiking neural networks to model EEG data [30]; an aim of this work is to produce a tool to explore and diagnose neurological disorders such as dementia – and not to use neuromodulation as a tool to aid artificial agents in achieving goals. Supervised learning methods and 'context-dependent plasticity' – or 'activational plasticity' [6] – have been shown to be beneficial for maintaining high accuracy for large numbers of sequential classification tasks, based on the MNIST and ImageNet datasets [31]; this was achieved by gating activations *randomly* in the network for each task. In other work, 'context-dependent selective activation' is achieved by learning parameters of a separate neuromodulatory network, which in turn gates activity for a prediction network [32]; this two-layered neural network approach is used for learning sequential tasks. This approach *indirectly* modulates learning, as the amount of activity in the predictive network after modulation is reflected in the back-propagation process.

Whilst it is common for learning and activity to be regulated by a separate group of modulatory neurons or an entire network [19], [20], [32], a distinguishing characteristic of our work is that we explore the impact that regulating activity-propagating signals *within a single neural network* has on an agent's ability to learn multi-stage tasks. By not explicitly regulating learning, we regulate behaviour to provoke immediate phenotypic changes based on environmental stimuli. Additionally, we use neuroevolution to evolve which neurons in the neural network are modulatory, resulting in a more structured way of operationalising neuromodulation than [31] for example, where neuronal activity is gated randomly.

D. Learning Multi-Stage Tasks in Multi-Agent Environments

Both humans and animals find learning in environments that change state or context without explicit cues challenging - this however is a characteristic of most realistic environments [1];

these changes need to be detected in order to adapt behaviour accordingly, as it is rare for this information to be explicitly available. This has also been identified as being a difficulty of learning multi-stage tasks, as the full state-space of tasks is not usually available when learning [17]; changes in state or stimuli also change the context in which behaviours are learnt.

These challenges are also present when neural networks learn to achieve new or many tasks, or navigate dynamic or uncertain environments; encoded knowledge must be adapted in order to learn new things [8].

Regulating synaptic plasticity with neuromodulation has been shown to facilitate adaptation and learning when there are changes in environmental stimuli or the task at hand, thus helping agents to overcome these issues [8], [16], [20], [25]. Whilst neuromodulation has also been used in multi-agent contexts, this is typically to explore the effect on cooperative or competitive strategies in social dilemmas [21] or in competitive environments [2], where agents are *explicitly* aware of others and thus employ strategies intentionally. Agents acting in novel environments may not have full or even partial information about others in the environment, and thus cannot cooperate or compete intentionally. In previous work, we show that learning in multi-agent environments without knowledge of the existence of others is problematic, as environmental stimuli change unpredictably as a result of the actions of others in the environment [11]. Social action is shown to improve learning in agents situated in multi-agent environments [11], however agents do not exhibit behavioural plasticity or use neuromodulation; furthermore, the study is limited to exploring multi-stage tasks.

In this paper, we aim to explore the challenges presented to neural controllers when they experience unpredictable environmental stimuli, and observe the effect that behavioural plasticity arising from the regulation of activity-propagating signals has on learning. As seen in the natural world [3], we would expect agents capable of behavioural plasticity to adapt better to changing, dynamic and uncertain environments, than those that are not. We investigate this by evolving agents to learn single- and multi-stage tasks, in both single- and multi-agent environments; this covers different combinations of environmental changes and variations. Specifically, we use the term 'multi-stage task' in a similar context to [17], where agents must learn multiple stages of a task in order to achieve a goal. Further, we explore the effect that changing the context in which an agent exists has on evolution, by evolving agents in an environment alone for a period of time and then evolving them for additional time within a multi-agent environment. We hypothesise that activational plasticity will help agents to achieve their tasks in these environments, by facilitating immediate behavioural changes in response to different environmental contexts or conditions.

III. Testbed and Agent Design

A. The River Crossing Dilemma Testbed

The River Crossing Dilemma (RCD) testbed was introduced by Barnes et al. [11], to explore how agents evolve to achieve individual goals in shared worlds; this extends the original River Crossing Task proposed by Robinson et al. [9]. Agents have no prior knowledge of the task or environment, and must learn what their goal is and how to achieve it without this information. The RCD is a 19×19 grid-world, with a two-cell deep river of Water in the centre. There are four Stones on each river bank, and all empty cells are Grass. An agent's goal is to collect one of its allocated Resources from either side of the river, rewarding the agent with a highly positive fitness. Conversely, stepping into the river causes the agent to drown, giving it a highly negative fitness. As a result, the task is multi-stage [17], as agents must evolve to perform sub-tasks and the appropriate behaviours that correspond to different states and environmental stimuli – they must build a bridge to cross the river to achieve their goal. As the river is two cells deep, two Stones must be placed in the same Water cell to successfully build a bridge; agents must step onto a cell with a Stone to pick it up, then stand adjacent to the river to partially or fully build a bridge. The RCD testbed is a bespoke Java implementation, and is presented in Figure 1. Time is measured in 'timesteps', where an agent can move a distance of one cell per timestep. For experiments with two agents, the agent that starts in the top left of the environment moves first, followed by the agent that starts in the bottom right.

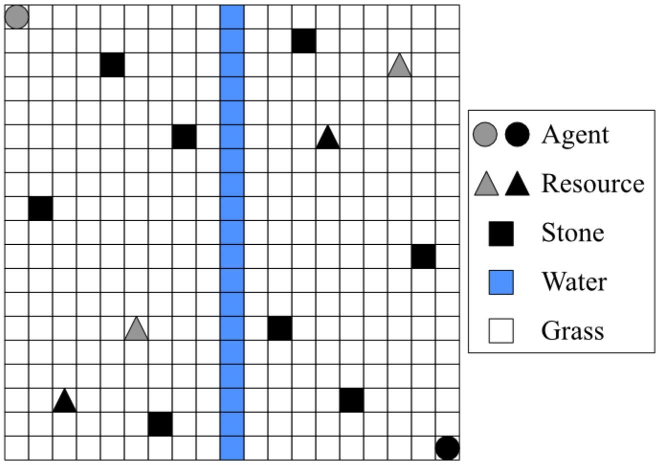


Fig. 1. The River Crossing Dilemma testbed, proposed by Barnes et al. [11]. The grey agent (top left) is allocated the two Resources in grey, and the black agent (bottom right) is allocated the two Resources in black; agents cannot interact with Resources not allocated to them. Both agents can interact with all other objects. For single-agent environments, the black agent is removed.

B. The Protected River Crossing Dilemma

We introduce the Protected River Crossing Dilemma (PRCD) – an adaptation of the RCD [11] specifically used to explore how agents evolve to solve single-stage tasks; like with the RCD, the PRCD is a bespoke Java implementation.

The environment is constructed as seen in Figure 1, but the river acts as an impassable – and most importantly, a non-lethal – obstacle, meaning agents cannot fall into it mistakenly. This simple change means that agents do not need to learn the different states in which they can interact with the river: that it is not safe unless the agent is carrying a Stone. As the PRCD river is impassable, agents must still perform sub-tasks such as bridge-building to succeed; removing the river entirely would remove the multi-stage task but also make the task trivial.

The single-stage task therefore reduces the variability in the task and environment, making it less complex; as plasticity is said to increase with environmental variability [33], we would expect the effect of neuromodulation to be less apparent than in the multi-stage RCD task. Further, we would expect that the benefit of neuromodulation is less evident still when agents evolve to solve the single-stage task alone compared to when they evolve together for this same reason.

C. Gamification of the RCD and PRCD

The RCD and PRCD are gamified, such that agents incur an increasing cost for each Stone placed in the river; a bridge is successfully built with two Stones. Therefore, in multi-agent environments, agents face a social dilemma and may either complete their task individually and be subjected to the full cost of bridge-building, cooperate to share the cost, or exploit other agents to avoid a cost at all. This creates a Snowdrift Game [34] (also known as the Chicken Game [21], [35]–[37] or the Hawk-Dove Game [21], [38]), resulting in less incentive to cooperate due to the cost of bridge-building, but failure for defection if the agent isn't able to achieve its goal. The fitness, or payoff, for agent p_i is calculated with Equation 1:

$$p_i \quad = \quad \frac{r_i}{N} \quad - \quad \left[\quad \frac{C \times s_i}{2}\Big(1 + s_i\Big) \quad \right] - f \qquad (1)$$

where r is the number of Resources collected by p_i, $N = 2$ and is the number of Resources allocated to each agent, $C = 0.1$ and is the cost of placing a Stone in the river, s_i is the number of Stones placed in the river by agent p_i, and $f = 1$ if agent p_i falls in the river, or 0 otherwise. An agent's fitness therefore records its own behaviour.

Commonly observed fitnesses are presented in a payoff matrix in Table I, using Equation 1. The maximum fitness an agent can achieve alone is 0.7, which increases to 0.9 if the cost of bridge-building is shared, or 1.0 if an agent exploits another in a shared environment; anything below 0.7 indicates the goal is not achieved.

TABLE I
PAYOFF MATRIX FOR THE RIVER CROSSING DILEMMA [11] AND PROTECTED RIVER CROSSING DILEMMA TESTBEDS, ASSUMING THAT BOTH RESOURCE OBJECTS HAVE BEEN RETRIEVED. S_x IS EQUAL TO THE NUMBER OF STONES PLACED BY AGENT x.

Agent 2 Agent 1	$S_2 = 0$	$S_2 = 1$	$S_2 = 2$
$S_1 = 0$	0.0 0.0	-0.1 0.0	0.7 1.0
$S_1 = 1$	0.0 -0.1	0.9 0.9	0.7 0.9
$S_1 = 2$	1.0 0.7	0.9 0.7	0.7 0.7

D. Agent Design

Agents in both the RCD and the PRCD use a two-layered neural network architecture, adapted from [11] and inspired by [9]. The deliberative layer generates high-level sub-goals based on the current inputs, corresponding to the agent's current state. This network is therefore responsible for the decision-making processes of agents; depending on the inputs and the weights of the network, the outputs indicate what the agent decides to do next in terms of sub-goals (whether it is attracted to, neutral towards or repulsed from certain objects in the environment). The weights of the network (as well as the type of each neuron in the network) therefore represent the genes of the agent, and therefore what behaviours it will exhibit depending on what inputs. The inputs are 1 or 0 depending on whether the agent is on Grass, a Resource, Water or a Stone, if it is currently carrying a Stone, and if a bridge has been built partially in the environment (i.e. one Stone in the river out of two). This fully-connected feed-forward network has six input neurons, three hidden layers with eight, six and four neurons respectively, and an output layer of three neurons (Figure 2). Resources, Stones and Water will be attractive if the output is 1, avoided if -1, or neutral if 0.

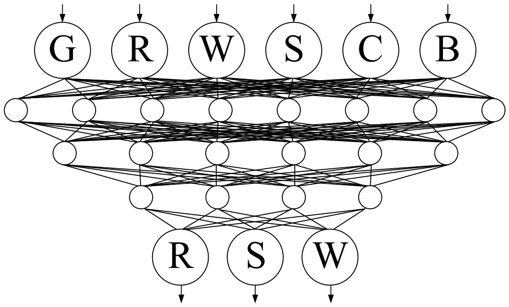


Fig. 2. *The Deliberative Layer* is a fully-connected neural network with three hidden layers, that generates high-level sub-goals. Inputs are 1 or 0, corresponding to the agent's current state: Grass, Resource, Water, Stone, Carrying Status, if a Bridge partially exists. Outputs are 1 for attraction, 0 for neutral or -1 for avoidance for each sub-goal: Resource, Stone, Water.

The reactive layer is a neural network with the same dimensions as the environment – in this case, 19×19 – where each neuron is connected to the surrounding eight neurons. This reactive neural network uses the shunting equation (Equation 2, [9], [11], [39], [40]) to create dynamic activity landscapes at each timestep based on the current sub-goals; agents can therefore hill-climb towards the goals generated in the previous layer by moving to the cell in its Moore neighbourhood (the surrounding eight cells) with the highest activity. Agents must make one move per timestep – they cannot remain stationary. Additionally, an agent will pick up a Stone automatically if it moves onto a cell with a Stone; an agent will also put a Stone in the river automatically if the cell to its left or right is Water – and if it is carrying a Stone. Equation 2 calculates the activity of each neuron based on its own and the surrounding activations: A is the passive decay rate; x_i is the current neuron; w_{ij} is the weight between neurons x_i and x_j, where x_j is one of the surrounding cells in x_i's Moore neighbourhood (indicated by $k = 8$); $[x_j]^+$ is calculated by $max(0, x_j)$, meaning that negative activity cannot propagate through the network. I is the Iota value of the neuron, which depends on the sub-goals from the deliberative layer (for a value of: 1, $I = 15$; -1, $I = -15$; and $I = 0$ otherwise); this creates hills and valleys in the activity landscape, as inspired by the original RCT testbed [9].

$$\frac{dx_i}{dt} = -Ax_i + I_i + \sum_{j=1}^{k} w_{ij}[x_j]^+ \tag{2}$$

E. Operationalising Activity-Gating Neuromodulation

Neuromodulated agents regulate their behaviour by gating activation *within* the neural network in the deliberative layer (Figure 2); this distinguishes our approach from others, which either use a separate modulatory network/neurons, or regulate learning as well as, or instead of, behaviour [19], [20], [32].

Figure 3 shows an example of this activity-gating modulation. Neurons in the deliberative layer may evolve to be non-modulatory or modulatory; if the incoming signal to a modulatory neuron is negative, it will fire and regulate behaviour by outputting a signal of 0 on each of its outgoing connections. This means that weights on the connections will effectively be 'turned-off', or gated, as the signal is blocked locally. Immediate, and more importantly reversible behavioural changes can therefore be achieved depending on the stimuli experienced. This gating or *modulation* of activity-propagating signals results in behavioural plasticity; an agent's genotype, represented by the evolved weights of the neural network and the types of the neurons in the deliberative layer, is therefore able to express multiple phenotypes depending on state and environmental stimuli – without changing, or potentially destroying, the knowledge encoded in the weights of the network. In other words, one deliberative neural network with modulatory neurons can exhibit temporary and reversible changes in behaviour depending on the stimuli and inputs; this is because modulatory neurons that are 'switched off' do not propagate any activity signals to the next layer of neurons, thus changing the output of the network and the resulting behaviour of the agent.

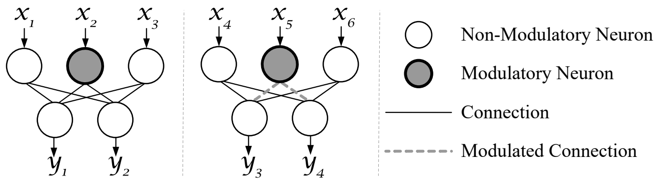


Fig. 3. *Activity-Gating Neuromodulation:* Modulatory neurons propagate activity the same as non-modulatory neurons when the incoming signal is $\geq$ 0; here, if the incoming activity signal to x_2 is positive, the outgoing activity signals of x_2 propagate as usual and are passed on to the next layer of neurons (in this case, y_1 and y_2). If, however, the incoming activity signal to x_5 is negative, the modulatory neuron fires and the outgoing activity is gated; specifically, this means that the neuron x_5 will output signals of 0 along each of its outgoing connections (in this case to y_3 and y_4), so the outgoing signal is effectively gated or 'turned off' when the signal is multiplied by the weight of the connection. This means agents can exhibit behavioural plasticity, as the weights of the neural network are not *changed*, but temporarily suppressed; this leads to the network producing different outputs and therefore different behaviours, without modifying the network weights in a permanent way. It is important to note that modulatory neurons only affect their own outgoing connections, so the connections from x_4 and x_6 to y_3 and y_4 are not affected when x_5 fires.

F. Evolutionary Algorithm

All experiments are conducted using the RCD testbed with the following common parameters, inspired by [11]. For each experiment, a population of 25 randomly initialised agents is evolved using a Steady State Genetic Algorithm. Agents acquire knowledge, and therefore 'learn', through evolution – there is no within-lifetime learning. At each generation, three agents are randomly selected from the population and are evaluated in a tournament; each agent has 500 timesteps to navigate the environment and achieve their goal. The evaluation stops if all agents reach the maximum amount of timesteps, achieve the goal, or die. The agent with the worst fitness in each tournament is replaced with an offspring generated from the best two. For each chromosome (layer of weights in the deliberative layer), this offspring has a probability of $P_{one} = 0.95$ to inherit the chromosome from a random parent, otherwise single-point crossover is used. Each connection weight w in the offspring's deliberative layer is then mutated by a random value from a Gaussian distribution with $\mu = w$ and $\sigma = 0.01$.

For neuromodulatory agents, the neurons in the hidden layers of the deliberative neural network are evolved in addition to the weights (input and output neurons cannot be modulatory); neurons may evolve to be standard non-modulatory neurons, or activity-gating modulatory neurons. The type of each neuron (modulatory or non-modulatory) is therefore not specified in advance, but evolved with neuroevolution like the weights of the network. The deliberative neural network of each agent is initialised with only non-modulatory neurons at the start of evolution. At each generation, the new offspring inherits the neuronal structure from a randomly chosen parent, where the parents are the two agents with the best fitnesses in the tournament as described above; there is a probability of $P_{mut} = 0.15$ that one randomly chosen neuron out of the three hidden layers in the deliberative network (Figure 2) will be mutated, from non-modulatory to modulatory or vice versa. This mutation rate is adapted from the mutation operators and probabilities used in [8]. Modulatory neurons regulate activity as outlined in Section III-E. Agents that do not use neuromodulation have a static network of non-modulatory neurons that do not evolve.

IV. Experimental Design

The experiments in this study aim to investigate the effect that behavioural plasticity through activity-gating neuromodulation has on agent evolution when the environment is prone to change; we use a series of experiments, outlined below, to explore the extent to which the ability to rapidly and reversibly change phenotypic behaviour helps agents to solve tasks in varying environmental conditions. All experiments are repeated 100 times using the same 100 seeds, both with and without neuromodulation. The first four sets of experiments evolve agents for 500,000 generations from a randomly-initialised state; the final set of experiments evolves agents for 1,000,000 generations in total by first evolving agents in an environment alone, and then continuing to evolve together.

The first set of experiments explore how agents evolve to solve a single-stage task in the Protected River Crossing Dilemma (PRCD), when they exist alone in the environment. This environment has the least inherent variability, which will provide a baseline to compare the effects of neuromodulation in the later experiments.

The second set of experiments introduces another agent into the single-stage task PRCD environment, which creates a social dilemma as agents may evolve to cooperate or exploit the other unintentionally. As agents cannot perceive or reason about the actions or existence of other agents, their environment appears unpredictable and therefore harder to evolve in. These experiments evolve two separate, randomly-initialised populations of agents that start on opposite corners of a shared PRCD world. In these experiments and the others that involve multi-agent environments, only the agent that begins in the top-left corner is assessed, so all results are comparable.

The third set of experiments investigates how agents that exist alone evolve to solve a multi-stage task, by instead using the RCD environment. This also adds an element of variability and uncertainty compared to the first set of experiments.

The fourth set of experiments use the RCD environment to explore how agents that share an environment together evolve to solve multi-stage tasks. Of these four experiments, this environment is the most variable, due to the imperceptible actions of the other agent within the environment, as well as the challenge of evolving to solve the multi-stage task. We expect to observe the most pronounced benefit of neuromodulation and behavioural activity in these experiments, as behavioural changes are increasingly useful as environmental conditions change [3].

The final set of experiments adds further unpredictability to the task; here, we explore how agents that have evolved alone for an initial period of 500,000 generations are able to achieve their goals and adapt when they continue to evolve in a shared environment with another agent for a further 500,000 generations. Agents therefore evolve for a total of 1,000,000 generations, with the initial period of evolving alone being identical to the first set of experiments. Here, we explore the extent to which activity-gating neuromodulation affects how agents evolve to solve multi-stage tasks, when the environment explicitly changes context from a single- to a multi-stage environment – this is the most challenging of the five sets of experiments due to the increase in environmental changes.

V. Results

A. Learning Single-Stage Tasks When Alone

We start by investigating how agents are able to evolve to solve the simplest task in the least variable environment in the study – the single-stage task in the Protected River Crossing Dilemma (PRCD) – and the role that neuromodulation plays.

Figure 4(a) shows the mean best-in-population fitness over time when agents evolve alone in the PRCD, both with and without neuromodulation. Here, neuromodulation appears to be increasingly beneficial over the course of evolution. Once the effect of neuromodulation is sustained, there is a clear benefit to behavioural plasticity in this environment once agents have finished evolving; 85% of agents were able to solve the single-stage task and achieve their goal with neuromodulation, compared to only 40% of agents that did not use neuromodulation (Table II).

B. Learning Single-Stage Tasks When Together

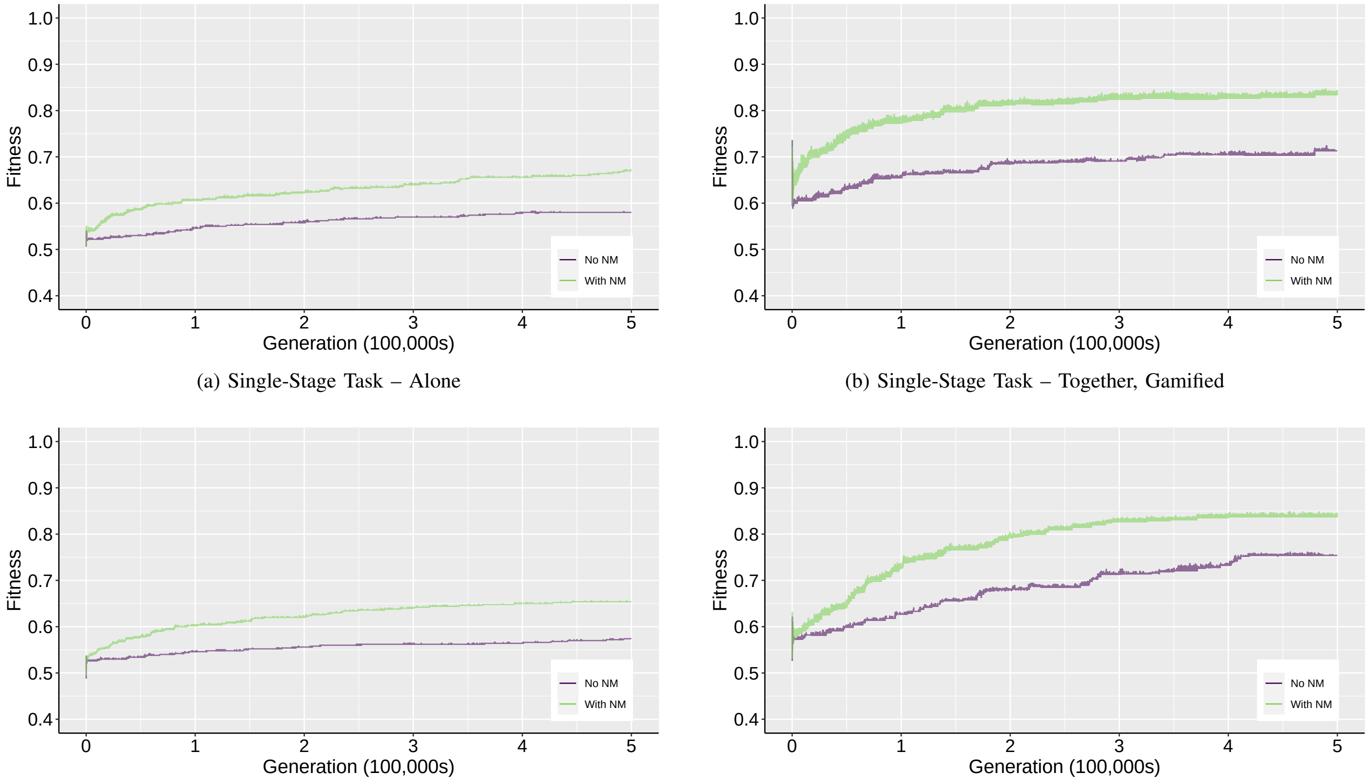
By introducing two agents into the single-task PRCD environment, the variability of the environment increases, and the task becomes gamified. Agents may evolve to achieve their goal alone, cooperate unintentionally, or exploit the actions of the other agent; agents therefore have the potential to achieve a higher fitness, at the risk of relying on the actions of another to achieve their goal. Agents are unable to perceive others or their actions, so the environment becomes unpredictable when it is shared with another agent.

Figure 4(b) shows the mean best-in-population fitness of agents evolving together in a shared PRCD environment. Similarly to when agents evolve to solve a single-stage task alone (Figure 4(a)), the effect of neuromodulation is slow to

TABLE II

The percentage of agents that receive common fitnesses in each experiment, after 500,000 generations of solving a single- (S) or multi- (M) stage task. 0.7 is a goal-achieving fitness after a bridge is built with two Stones; 0.9 is sharing the cost of bridge-building; 1.0 is exploitation; < 0.7 does not achieve the goal; ≥ 0.7 is a goal-achieving fitness.

Experiment	Task	Fitness (% of Agents)				
	(S/M)	0.7	0.9	1.0	< 0.7	≥ 0.7
Alone	S	40	0	0	60	40
Alone with NM	S	85	0	0	15	85
Alone	M	37	0	0	63	37
Alone with NM	M	77	0	0	23	77
Together	S	29	5	27	39	61
Together with NM	S	49	2	46	3	97
Together	M	27	5	36	32	68
Together with NM	M	44	0	50	6	94
CE	M	40	1	32	27	73
CE with NM	M	47	2	50	1	99



(a) Single-Stage Task – Alone

(b) Single-Stage Task – Together, Gamified

(c) Multi-Stage Task – Alone

(d) Multi-Stage Task – Together, Gamified

Fig. 4. The mean best-in-population fitnesses of agents evolving to solve (a) a single-stage task alone, (b) a single-stage task together, (c) a multi-stage task alone and (d) a multi-stage task together, for 500,000 generations, with and without neuromodulation (NM). Single-stage tasks take place in the PRCD, and multi-stage tasks take place in the RCD. A fitness of: 0.7 indicates the goal is achieved individually; 0.9 indicates the cost of bridge-building is shared; 1.0 indicates an agent exploits another's act of building a bridge; 0.7 or above indicates the goal is achieved; below 0.7 indicates the task is failed (Equation 1).

manifest; however, as agents can access a higher fitness than they can achieve alone, the effect of neuromodulation is more prominent than when agents are alone. In Figure 4(b), agents evolve to achieve a higher fitness more often, and by the end of evolution, 97% of neuromodulatory agents achieve their goal compared to 61% of non-modulatory agents.

C. *Learning Multi-Stage Tasks When Alone*

The multi-stage task present in the RCD creates a more variable environment than that seen in the single-stage task PRCD environment; agents must evolve to match correct behaviours with different environmental stimuli under different conditions, which is a more challenging – and more perilous – task when the possibility of falling in the river exists.

When agents evolve alone in the RCD environment, they can only achieve their goal once they have built a bridge on their own. As the environment is gamified, the maximum fitness an agent can achieve is therefore 0.7, after the bridge-building cost is deducted from the fitness (Equation 1).

The mean best-in-population fitness increases over time as more agents evolve successful solutions; after 500,000 generations, 37% of agents evolved the necessary behaviours to achieve their goal without neuromodulation, compared to 77% that achieved their goal with neuromodulation (Table II). Figure 4(c) shows that the mean best-in-population fitness increases faster when agents use neuromodulation, indicating that agents are more likely to evolve successful solutions, and that they are able to do this in fewer generations than agents that do not use neuromodulation. The increase in task complexity and environmental variability compared to the single-stage task PRCD environment indicates that behavioural plasticity is more beneficial when there is greater variability or uncertainty in the environment or the task at hand.

D. *Learning Multi-Stage Tasks When Together*

The fitness function presented in Equation 1 evaluates each agent individually. When agents share an environment, they can still achieve their goal alone by building a bridge completely by themselves and enduring the associated cost, but they can also exploit the other to avoid the cost, or cooperate and share the cost of bridge-building. The maximum accessible fitness to each agent therefore increases to 1.0 instead of 0.7, as agents may achieve their goal without building a bridge. In each case, agents have no capacity to perceive the existence or actions of the other, so cannot cooperate or exploit intentionally; instead, these agents perceive changes in environmental stimuli, and attempt to adapt their behaviour accordingly.

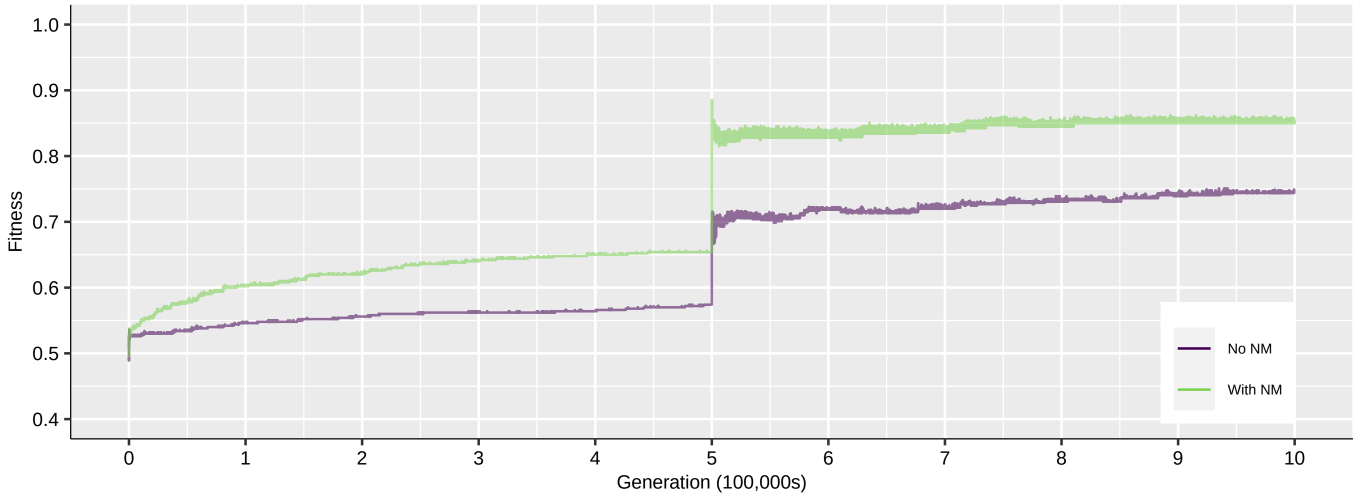


Fig. 5. The mean best-in-population fitness of agents that evolve alone for 500,000 generations, then continue to evolve together with a random partner for a further 500,000 generations, with and without neuromodulation (NM). A fitness of 0.7 or above indicates the goal is achieved (Equation 1).

The multi-stage task presented in the RCD adds yet another layer of complexity onto the task and the environment; a multi-agent environment introduces an element of unpredictability as agents cannot perceive others, and a multi-stage task means that the agent must discover multiple states and the corresponding consequences in the environment in order to achieve its task.

Figure 4(d) shows that neuromodulated agents that regulate their behaviour evolve to achieve their goal more often, and in fewer generations, than non-modulated agents. After 500,000 generations, 94% of neuromodulated agents achieve their goal, compared to only 68% of non-modulated agents (Table II). This shows that agents receive a benefit from expressing behavioural plasticity in response to changes in environmental stimuli caused by the actions of others.

E. Learning a Multi-Stage Task with Continued Evolution

When agents evolve alone in the RCD, the maximum fitness they can achieve is 0.7 after the total cost of building a bridge is deducted. When agents evolve together, this threshold increases to 1.0 as the possibility to exploit the bridge-building of other agents arises. In the following experiments, agents undergo an initial period of evolution in the multi-stage RCD environment alone for 500,000 generations. After this initial period of evolution, agents are then paired with another agent who has also evolved alone, and both continue to evolve together in a shared, multi-stage task RCD environment for a further 500,000 generations. By changing the agents' environment from an individual to a shared environment, the predictability decreases not only because the environment is now shared – but because the context in which the agents have evolved in is completely changed. Agents must adapt their behaviour to cope with a change in environmental stimuli, and the unanticipated actions of others in the environment.

Figure 5 shows the full evolution of the agents that evolve for a period alone (generations 1-500,000), and then continue to evolve in a shared environment with another agent (generations 500,001-1,000,000). The change in context from a single-agent to a multi-agent environment has an instantaneous effect on the evolution of agents, which can be seen in the sharp increase in fitness around generation 500,000. This is because agents can immediately capitalise on the changes in environmental stimuli caused by the imperceptible actions of the other agent in the environment – both with and without neuromodulation. Neuromodulation is observed to help agents to solve the multi-stage task in the RCD when they are alone, and continue to help them adapt to their new, shared environment when the context of the task is changed. The benefit of neuromodulation is maintained for the remainder of the evolutionary process, resulting in 99% of agents achieving their goal, compared to only 73% of non-modulatory agents.

VI. Discussion and Further Analysis

Activity-gating neuromodulation appears to increase both the likelihood and the speed that agents evolve successful solutions – both when they exist alone, and when they exist together (Figure 4). This observation is more prominent when agents evolve to solve a multi-stage task compared to a single-stage task, thus showing that a simple change in task complexity can affect the expected fitness of agents and their ability to achieve their goals.

Similarly, a more obvious benefit arising from the use of neuromodulation is seen in agents that evolve in multi-agent environments than in those that evolve alone. This is because the actions of each agent change the context of the environment and therefore the state of each agent within it, causing the variability of the environment to increase. The benefit arising from a capacity for behavioural plasticity through neuromodulation is therefore observed to increase as the variability and unpredictability of the task and environment increases. Agents that evolve to solve multi-stage tasks in multi-agent environments are observed to receive the highest benefit from immediate and reversible behavioural changes in response to environmental stimuli. Phenotypic plasticity is said to promote better adaptation to variable and changing environmental conditions [3], which is seen when evolving to solve multi-stage tasks in multi-agent environments; these are the most dynamic and uncertain conditions in this study.

The mean, median and variance for the best-in-population fitness of each experiment was calculated after agents had evolved for 500,000 generations, presented in Table III. This analysis shows that neuromodulatory agents can expect a higher median and mean fitness across all experiments; the exception to this is when agents evolve to solve a single-stage task together, in which case the median fitness is the same with and without neuromodulation. Combined with the results presented in Table II, neuromodulatory agents can therefore not only be expected to have a higher mean and median fitness, but they can be expected to solve the task and achieve their goal more often than non-modulatory agents; this is observed both in single- and multi-stage tasks, as well as single- and multi-agent environments. The variance in the best-in-population fitness after evolution is also lower in neuromodulatory agents, which further exemplifies the benefits of behavioural plasticity.

To analyse the effect that activity-gating neuromodulation has on evolution further, Wilcoxon Signed Rank statistical tests were conducted to compare the best-in-population fitnesses of each experiment when agents evolve both without and with neuromodulation. This is a non-parametric test used to compare the medians of two paired distributions; the null hypothesis of a two-tailed test is that the distribution medians are equal, whereas one-tailed tests have the alternative hypothesis that there is a directional difference in the distribution medians (e.g. $m_n > m_m$). The null hypothesis can be rejected when the calculated p-value is significant, below 0.05. These results are presented in Table IV. The two-tailed tests show that there is a significant difference in median received fitness between non-modulated and neuromodulated agents, for each experiment in the study; the null hypothesis that the medians of the two distributions are equal, can thus be rejected as $p < 0.05$. Additionally, two one-tailed tests were conducted to investigate whether there was a directional difference in the distribution medians. These tests indicate that there is a significant directional difference in the medians of the two distributions, where the median of the non-modulatory approach (m_n) is lower than the modulatory approach (m_m) for each experiment conducted; furthermore, the contrasting one-tailed test ($m_n > m_m$) shows no significant difference. These results demonstrate that neuromodulation has both a significantly different and a positive effect on the expected fitness of agents, in all areas of the study.

TABLE III
MEAN, MEDIAN AND VARIANCE OF THE BEST-IN-POPULATION FITNESS AFTER $500{,}000$ GENERATIONS OF EVOLVING ALONE, TOGETHER, AND WITH CONTINUED EVOLUTION (CE) AFTER AN INITIAL PERIOD OF EVOLVING ALONE. THE HIGHEST MEAN AND MEDIAN, AND LOWEST VARIANCE ARE IN **BOLD**.

Experiment	Task	Mean	Median	Variance
Alone	S	0.58	0.5	0.00969697
Alone with NM	S	**0.67**	**0.7**	**0.005151515**
Alone	M	0.574	0.5	0.009418182
Alone with NM	M	**0.654**	**0.7**	**0.007155556**
Together	S	0.713	0.7	0.04215253
Together with NM	S	**0.836**	0.7	**0.02515556**
Together	M	0.754	0.7	0.04473131
Together with NM	M	**0.838**	**0.85**	**0.02864242**
CE	M	0.744	0.7	0.03844848
CE with NM	M	**0.852**	**0.95**	**0.02332929**

TABLE IV
WILCOXON SIGNED RANK TESTS STATISTICAL TESTS, PAIRING THE BEST-IN-POPULATION FITNESS OF THE NON-MODULATORY (m_n) AND MODULATORY (m_m) APPROACHES FOR EACH EXPERIMENT: EVOLVING IN AN ENVIRONMENT ALONE, TOGETHER, OR WITH CONTINUED EVOLUTION (CE), WITH A SINGLE- (S) OR MULTI- (M) STAGE TASK. SIGNIFICANT p-VALUES AT THE $p < 0.05$ LEVEL ARE **BOLD**

Exp	Task	Statistical Test Alternative Hypothesis		
	(S/M)	$m_n \neq m_m$	$m_n < m_m$	$m_n > m_m$
Alone	S	**0.000000002588**	**0.000000001294**	1
Together	S	**0.0002362**	**0.0001181**	0.9999
Alone	M	**0.00000002994**	**0.00000001497**	1
Together	M	**0.01594**	**0.00797**	0.9922
CE	M	**0.0002593**	**0.0001296**	0.9999

VII. Conclusion

Increasing environmental variability makes learning challenging for neural controllers, as encoded information must be overwritten in order to learn new things when environmental conditions change. The capacity to immediately and reversibly change behaviour based on environmental stimuli is said to promote adaptation in variable environments [3], [6]. We have thus investigated the effect that activity-gating neuromodulation has on an agent's ability to evolve to succeed and to make decisions in environments of increasing variability, by exploring both single- and multi-stage tasks in single- and multi-agent environments.

This study uses the River Crossing Dilemma [11] to explore how agents evolve to solve multi-stage tasks; additionally we propose a new adaptation of the testbed called the Protected River Crossing Dilemma, in order to observe how agents evolve to solve simpler single-stage tasks in less variable environments. Our results demonstrate that behavioural plasticity as a result of activity-gating neuromodulation has a significant and increasingly positive effect on the expected fitness of evolved agents, when the variability of the environment increases; this behavioural plasticity is beneficial to create adaptive agent controllers that can temporarily and reversibly change behaviour in novel environments or situations.

We also show that when the context of an agent's environment changes from being individual to shared, neuromodulation helps agents to adapt and succeed to the new context and change in environmental stimuli. Often, agents in this study will evolve an 'exploitative' fitness, meaning that their success – and higher fitness – relies on the actions of others in the environment; future work will explore the extent to which behavioural plasticity enables agents to maintain their goal-achieving behaviours when the presence of other agents in the environment is unpredictable and uncertain.

Acknowledgements

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 262762. The authors would like to thank Aston University and the University of Oslo for supporting the research visits during this collaboration.

References

[1] T. Qian, T. F. Jaeger, and R. Aslin, "Learning to represent a multi-context environment: More than detecting changes," *Frontiers in Psychology*, vol. 3, p. 228, 2012. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2012.00228

[2] J. Yoder and L. Yaeger, "Evaluating topological models of neuromodulation in polyworld," in *ALIFE 14: Proceedings of The Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, no. 26, 2014, pp. 916–923.

[3] M. Viney and A. Diaz, "Phenotypic plasticity in nematodes," *Worm*, vol. 1, no. 2, pp. 98–106, 2012, pMID: 24058831. [Online]. Available: https://doi.org/10.4161/worm.21086

[4] R. F. Oliveira, "Social plasticity in fish: integrating mechanisms and function," *Journal of Fish Biology*, vol. 81, no. 7, pp. 2127–2150, 2012.

[5] T. L. Rymer, N. Pillay, and C. Schradin, "Extinction or survival? behavioral flexibility in response to environmental change in the african striped mouse rhabdomys," *Sustainability*, vol. 5, no. 1, pp. 163–186, 2013.

[6] E. C. Snell-Rood, "An overview of the evolutionary causes and consequences of behavioural plasticity," *Animal Behaviour*, 2013.

[7] K. O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen, "Designing neural networks through neuroevolution," *Nature Machine Intelligence*, 2019.

[8] K. O. Ellefsen, J. B. Mouret, and J. Clune, "Neural Modularity Helps Organisms Evolve to Learn New Skills without Forgetting Old Skills," *PLoS Computational Biology*, 2015.

[9] E. Robinson, T. Ellis, and A. Channon, "Neuroevolution of agents capable of reactive and deliberative behaviours in novel and dynamic environments," in *Advances in Artificial Life*. Springer, 2007, pp. 1–10.

[10] J. Borg, A. Channon, and C. Day, "Discovering and Maintaining Behaviours Inaccessible to Incremental Genetic Evolution Through Transcription Errors and Cultural Transmission," in *ECAL*, 2011.

[11] C. M. Barnes, A. Ekárt, and P. R. Lewis, "Social action in socially situated agents," in *Proceedings of the IEEE 13th International Conference on Self-Adaptive and Self-Organizing Systems*, 2019, pp. 97–106.

[12] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423–1447, Sep. 1999.

[13] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, 2002.

[14] M. McCloskey and N. J. Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem," *Psychology of Learning and Motivation - Advances in Research and Theory*, 1989.

[15] J. A. Bullinaria, "Understanding the emergence of modularity in neural systems," *Cognitive Science*, 2007.

[16] R. Velez and J. Clune, "Diffusion-based neuromodulation can eliminate catastrophic forgetting in simple neural networks," *PLoS ONE*, 2017.

[17] A. Dezfouli and B. W. Balleine, "Learning the structure of the world: The adaptive nature of state-space and action representations in multi-stage decision-making," *PLOS Computational Biology*, vol. 15, no. 9, pp. 1–22, 09 2019.

[18] L. F. Abbott, "Modulation of function and gated learning in a network memory," *Proceedings of the National Academy of Sciences of the United States of America*, 1990.

[19] N. Vecoven, D. Ernst, A. Wehenkel, and G. Drion, "Introducing neuromodulation in deep neural networks to learn adaptive behaviours," *PLOS ONE*, vol. 15, no. 1, pp. 1–13, 01 2020.

[20] A. R. Daram, D. Kudithipudi, and A. Yanguas-Gil, "Task-based neuromodulation architecture for lifelong learning," in *20th International Symposium on Quality Electronic Design (ISQED)*, 2019, pp. 191–197.

[21] D. E. Asher, A. Zaldivar, B. Barton, A. A. Brewer, and J. L. Krichmar, "Reciprocity and retaliation in social games with adaptive agents," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 3, pp. 226–238, 2012.

[22] F. Mery and J. G. Burns, "Behavioural plasticity: An interaction between evolution and experience," *Evolutionary Ecology*, 2010.

[23] J. A. Stamps, "Individual differences in behavioural plasticities," *Biological Reviews*, 2016.

[24] A. W. Hamood and E. Marder, "Animal-to-animal variability in neuromodulation and circuit function," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 79. Cold Spring Harbor Laboratory Press, 2014, pp. 21–28.

[25] A. Soltoggio, J. A. Bullinaria, C. Mattiussi, P. Dürr, and D. Floreano, "Evolutionary advantages of neuromodulated plasticity in dynamic, reward-based scenarios," in *Artificial Life XI: Proceedings of the 11th International Conference on the Simulation and Synthesis of Living Systems, ALIFE 2008*, 2008.

[26] L. F. Abbott and S. B. Nelson, "Synaptic plasticity: taming the beast," *Nature Neuroscience*, vol. 3, no. 11, pp. 1178–1183, 2000.

[27] W. S. Grant, J. Tanner, and L. Itti, "Biologically plausible learning in neural networks with modulatory feedback," *Neural Networks*, 2017.

[28] J. Huang, X. Ruan, N. Yu, Q. Fan, J. Li, and J. Cai, "A cognitive model based on neuromodulated plasticity," *Computational Intelligence and Neuroscience*, 2016.

[29] J. L. Krichmar, "The neuromodulatory system: A framework for survival and adaptive behavior in a challenging world," *Adaptive Behavior*, vol. 16, no. 6, pp. 385–399, 2008.

[30] J. I. Espinosa-Ramos, E. Capecci, and N. Kasabov, "A computational model of neuroreceptor-dependent plasticity (nrdp) based on spiking neural networks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 1, pp. 63–72, March 2019.

[31] N. Y. Masse, G. D. Grant, and D. J. Freedman, "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. E10 467–E10 475, 2018.

[32] S. Beaulieu, L. Frati, T. Miconi, J. Lehman, K. O. Stanley, J. Clune, and N. Cheney, "Learning to Continually Learn," *arXiv e-prints*, p. arXiv:2002.09571, Feb. 2020.

[33] P. E. Komers, "Behavioural plasticity in variable environments," *Canadian Journal of Zoology*, vol. 75, no. 2, pp. 161–169, 1997.

[34] K. Mogielski and T. Płatkowski, "A mechanism of dynamical interactions for two-person social dilemmas," *Journal of Theoretical Biology*, 2009.

[35] P. Kollock, "Social Dilemmas: The Anatomy of Cooperation," *Annual Review of Sociology*, 1998.

[36] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent Reinforcement Learning in Sequential Social Dilemmas," *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.

[37] P. A. van Lange, J. Joireman, C. D. Parks, and E. Van Dijk, "The psychology of social dilemmas: A review," *Organizational Behavior and Human Decision Processes*, 2013.

[38] K. Sigmund and M. A. Nowak, "Evolutionary game theory," *Current Biology*, 1999.

[39] S. X. Yang and M. Meng, "An efficient neural network method for real-time motion planning with safety consideration," *Robotics and Autonomous Systems*, vol. 32, pp. 115–128, 2000.

[40] ——, "An efficient neural network approach to dynamic robot motion planning," *Neural Networks*, vol. 13, no. 2, pp. 143–148, 2000.