Behavioural Plasticity Can Help Evolving Agents in Dynamic Environments But at the Cost of Volatility

CHLOE M. BARNES, Aston University, United Kingdom
ANIKÓ EKÁRT, Aston University, United Kingdom
KAI OLAV ELLEFSEN, University of Oslo, Norway
KYRRE GLETTE, University of Oslo, Norway
PETER R. LEWIS, Ontario Tech University, Canada
JIM TØRRESEN, University of Oslo, Norway

Neural networks have been widely used in agent learning architectures; however, learnings for one task might nullify learnings for another. Behavioural plasticity enables humans and animals alike to respond to environmental changes without degrading learnt knowledge; this can be achieved by regulating behaviour with neuromodulation – a biological process found in the brain. We demonstrate that by modulating activity-propagating signals, neurally trained agents evolving to solve tasks in dynamic environments that are prone to change can expect a significantly higher fitness than non-modulatory agents, and also achieve their goals more often. Further, we show that while behavioural plasticity can help agents to achieve goals in these variable environments, this ability to overcome environmental changes with greater success comes at the cost of highly volatile evolution.

CCS Concepts: • **Computing methodologies** → **Multi-agent systems**; *Neural networks*; *Artificial life*; • **Theory of computation** → *Evolutionary algorithms.*

Additional Key Words and Phrases: Neuroevolution, Neuromodulation, Behavioural Plasticity, Evolutionary Multi-agent Systems

ACM Reference Format:
Chloe M. Barnes, Anikó Ekárt, Kai Olav Ellefsen, Kyrre Glette, Peter R. Lewis, and Jim Tørresen. 2022. Behavioural Plasticity Can Help Evolving Agents in Dynamic Environments But at the Cost of Volatility. *ACM Trans. Autonom. Adapt. Syst.* 1, 1 (January 2022), 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Natural and artificial environments are often complex, unpredictable and dynamic, making learning and surviving a challenge for animals and artificial agents alike [32, 48]. In order to survive in these challenging conditions, many organisms, such as nematodes [43] and fish [31], show behavioural plasticity to rapidly adapt to novel situations by temporarily changing behaviour [34, 35]. This problem is not just specific to natural beings; artificial neural networks

Authors' addresses: Chloe M. Barnes, c.barnes1@aston.ac.uk, Aston University, Birmingham, United Kingdom, B4 7ET; Anikó Ekárt, a.ekart@aston.ac.uk, Aston University, Birmingham, United Kingdom, B4 7ET; Kai Olav Ellefsen, kaiolae@ifi.uio.no, Department of Informatics, University of Oslo, Oslo, Norway, NO-0316; Kyrre Glette, kyrrehg@ifi.uio.no, RITMO, Department of Informatics, University of Oslo, Oslo, Norway, NO-0316; Peter R. Lewis, peter.lewis@ontariotechu.ca, Ontario Tech University, Oshawa, Canada, ON L1G 0C5; Jim Tørresen, jimtoer@ifi.uio.no, RITMO, Department of Informatics, University of Oslo, Oslo, Norway, NO-0316.

(ANNs) for example are also often tasked with learning in dynamic and unpredictable environments. Encoding new information in ANNs can result in a degradation of performance and catastrophic forgetting when learning new tasks or experiencing novel environmental contexts [10, 13, 27, 42]; learnt knowledge must be changed to learn new things, often leading to knowledge loss [13]. One way ANNs can 'learn' is with neuroevolution, where ANNs are evolved with an evolutionary algorithm in accordance to a fitness function [38]. Many applications of neuroevolution focus on evolving the connection weights of ANNs [5, 9, 13, 33], however more complex approaches that evolve both the weights and topologies of ANNs exist [39, 46]. Learning complex, sequential or multi-stage tasks is often hard for these neural controllers, as complete information about the environment – including the available actions, their cues and their consequences – is not usually accessible [12, 32]; this is also evident when environments are shared, as the actions of individuals change the context of the environment for others [5]. In nature, behavioural plasticity can be achieved with *neuromodulation* – a biological process whereby chemical signals are regulated (often also termed 'modulated' or 'gated') in the brain depending on environmental stimuli [1]. Consequently, neuromodulation has been used to aid neural controllers with learning new or sequential tasks, and learning in dynamic environments [11, 13, 41].

We present an experimental study that abstracts these concepts to explore how neural controllers evolve to achieve goals in variable and dynamic environments, when they have no knowledge of the task, environment, or others. We then present a comprehensive analysis of the effect that behavioural plasticity has on evolution as an extension of previous work [4]; we ascertain whether neuromodulation affects the ability to achieve goals at the end of, and during, evolution, and ultimately whether this affects evolutionary volatility. Observing evolutionary volatility indicates how often fitness is prone to fluctuate during evolution – more volatility means less predictability in evolution. This is important to consider when designing systems – especially those in highly variable or dynamic environments – as a trade-off between fitness and predictability may need to be considered.

The experiments use the River Crossing Dilemma testbed [5], which is designed to explore social concepts of arbitrary complexity; as such, the study observes how agents evolve in single- and multi-agent environments. ANNs are just one example of an agent controller in which behaviour can be learnt; we use ANNs in line with previous River Crossing testbeds [5, 9, 33], to explore how ANNs make decisions in social environments to solve tasks of variable complexity. Here, we define a multi-stage task as one that an agent must learn, and pass, through multiple states, and perform different behaviours in different contexts to achieve their goal; this definition is inspired by [12].

We hypothesise that reversible and immediate behavioural changes as a result of neuromodulation will enable agents to overcome the challenges associated with solving tasks and achieving goals in unpredictable and dynamic environments with greater effect than non-plastic agents. We demonstrate this using the River Crossing Dilemma (RCD) testbed introduced by Barnes et al. [5] for multi-stage tasks, as well as a new adaptation of the testbed called the *Protected* River Crossing Dilemma (PRCD) [4] for exploring single-stage tasks. The effect that behavioural plasticity has on evolution can thus be observed in agents that evolve in different contexts and conditions. We operationalise neuromodulation by gating (regulating) activity *within* a single neural network, allowing agents to regulate their behaviour without affecting encoded knowledge; this distinguishes our approach from others, which either use a separate modulatory network/neurons, or regulate learning as well as, or instead of, behaviour [8, 11, 41]. By doing this, fewer resources are required for plastic behaviour – which becomes more critical as the size or complexity of the network increases. Further, we investigate how regulating behaviour may help agents to evolve in multi-agent environments, without the capacity to learn of the existence of others; 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 [3, 48], however our work extends the notion of Barnes et al. [5], where cooperation and exploitation may emerge but cannot be intended. A novelty of this study is therefore the exploration of how neuromodulation affects agent evolution when agents are unable to perceive the actions of others in the environment; we specifically structure the study around exploring how agents evolve to solve single- and multi-stage tasks, in single- and multi-agent environments. Additionally, we analyse the fitnesses that agents receive during and after evolution, as well as the evolutionary volatility experienced. These experiments are therefore designed to investigate the extent to which behavioural plasticity affects agent evolution and the ability to achieve goals.

2 BACKGROUND

2.1 Behavioural Plasticity and Neuromodulation

One way to design adaptive systems is by utilising behavioural plasticity; this can be seen as the ability to change or adapt behaviour based on changes in stimuli [28]. This is important for navigating uncertain, novel or dynamic environments and can be classed into two different types: developmental and activational [35]. 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. These behavioural changes are reversible, as the genotype remains unchanged. Activational plasticity is also termed 'innate' [28] or 'contextual' [37] plasticity.

Neuromodulation is a biological process found in animal brains [18], 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 [1, 13, 36]. In neuroscience, synaptic plasticity is the modification of synapses between neurons through strengthening or weakening them [2]. In ANNs, synaptic plasticity is achieved by modulating neural network weights. Developmental plasticity is thus 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, without affecting learning or synaptic strengths.

2.2 Achieving Developmental Plasticity with Neuromodulation

Similarly to ANNs being inspired by the connectionist architectures found in brains, neuromodulation has been widely applied to artificial models to regulate synaptic plasticity and the learning rate of neural connections. ANNs have been evolved with modulatory neurons to regulate learning and mitigate the catastrophic forgetting associated with performing tasks in uncertain environments [36]; this improves learning when agents forage in T-maze problems (either moving left or right, in a maze with a 'T'-shape), in which the location of the reward can change. Other studies show that promoting the evolution of modular neural networks by introducing a cost for neural connections can mitigate catastrophic forgetting and improve learning – which is regulated with neuromodulation [13]. Neuromodulation has also been used to develop conflict learning in ANNs [16], and associative learning in robots [20]; these two approaches employ neuromodulation, but do not use neuroevolution as a learning mechanism.

These approaches modulate learning, resulting in 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. Further, these approaches only use neuromodulation in ANNs or robots that exist in isolation; we however explore how immediate behavioural plasticity can be achieved with neuromodulation without regulating learning, in both single- and multi-agent environments.

2.3 Achieving Activational Plasticity with Neuromodulation

Neurobiological mechanisms have been explored using a computational framework based on neuromodulatory systems such as the dopaminergic and serotonergic systems, by regulating synaptic activity [24]. 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 explore biological systems computationally. The effects of modulating neuroreceptors and synaptic plasticity have been studied with spiking neural networks to model EEG data [14]; an aim of that work is to produce a tool to diagnose neurological disorders such as dementia – and not to use neuromodulation to aid artificial agents in achieving goals. Supervised learning methods and 'context-dependent plasticity' ('activational plasticity' [35]) have been shown to be beneficial for maintaining high accuracy for large numbers of sequential classification tasks, based on the MNIST and ImageNet datasets [26]; this was achieved by regulating activity *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 [8]. This two-tiered neural network approach is used for learning sequential tasks and *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 [8, 11, 41], 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 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 [26] for example, where neuronal activity is gated randomly.

2.4 Multi-task Reinforcement Learning

The vast majority of Reinforcement Learning problems are single-task, meaning an agent is trained to perform a specific task, such as playing a certain video game, and after training tested on the same task. Impressive progress and success has been demonstrated in single-task Reinforcement Learning in recent years due to the power of Deep Learning [25, 29]. Learning multiple tasks is a more challenging problem, because as new tasks are learned, competency on previous tasks needs to be retained. A straightforward way to fix this is to store all the previous training data in a replay buffer, mix them together, and thus train on all tasks at once. While efficient, this is not biologically plausible and does not scale to very large collections of tasks. A more realistic setting is one where learning has to happen in a sequence. Humans and animals learn things sequentially all the time [7], but AI systems tend to struggle with catastrophic forgetting of previously learned tasks when they do so [27]. Many approaches have been suggested for reducing the impact of such forgetting [13, 22, 40], but a general solution to this problem does not yet exist.

One particular solution that is closely related to the work in this paper, is to reduce catastrophic forgetting with neuromodulation [42]. That work evolved neural networks with modulatory signals that modified learning rates, demonstrating that the neuromodulation helped networks decompose the problem into subtasks, reducing forgetting and interference.

As mentioned above, we are in the current study interested in the less-studied activational plasticity, rather than long-term regulation of learning as applied in previous studies applying neuromodulation as a tool to reduce catastrophic forgetting. Rather than decomposing the multi-task learning problem, activational plasticity has the potential to learn everything together, and on-demand modify the ANN function depending on context.

2.5 Meta-Learning

A problem closely related to multi-task learning, and also closely related to the problem studied in this paper, is meta-learning. Meta-learning is the challenge of "learning to learn", that is, to optimize a model to as quickly as possible master new tasks that it is presented with. Those new tasks are never encountered during the initial optimization, forcing the agent to acquire general learning strategies [15].

Recent years have seen a large interest and exciting progress in this challenging problem. Influential contributions include MAML which optimizes weights of a neural network so that a few updates to those weights allow many new tasks to be mastered quickly [15], OML which learns intermediate representations that are frozen and later shared across many learned tasks [21], and ANML which trains a neuromodulated activity-gating ANN that protects from catastrophic forgetting during meta-learning [8].

The latter, which like our method relies on activity-gating neuromodulation, is the most closely related to this work. However, in addition to important differences in the algorithm and training procedure, ANML was focused on meta-learning for classification problems - whereas we here focus on evolving agents in multi-agent worlds.

2.6 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 [32], and information about these changes is rarely explicitly available. Learning multi-stage tasks is also difficult, as the full state-space of tasks is not usually available when learning [12]; changes in state or stimuli also change the context in which behaviours are learnt.

Navigating dynamic or uncertain environments, or learning to achieve new or many tasks, is challenging for ANNs; encoded knowledge must be adapted in order to learn new things [13]. Regulating synaptic plasticity with neuromodulation can facilitate adaptation and learning when the task or environment changes, thus helping agents to overcome these issues [11, 13, 36, 42]. Whilst neuromodulation has been used in multi-agent contexts, this is typically to explore the effect on cooperative or competitive strategies in social dilemmas [3] or in competitive environments [48], where agents are *explicitly* aware of others. Agents in novel environments may not have full or even partial information about others, and thus cannot cooperate or compete intentionally. In previous work, we have shown that learning in multi-agent environments without knowledge of others is problematic, as the actions of others change the environment unpredictably [5]. Social action is shown to improve learning in multi-agent environments [5], however the agents in that study do not exhibit behavioural plasticity; furthermore, the study is limited to exploring multi-stage tasks.

We aim to explore the challenges presented to neural controllers when they experience unpredictable environments, and observe the effect that behavioural plasticity arising from regulating activity-propagating signals has on evolution. As seen in the natural world [43], we would expect plastic agents to adapt better to changing and uncertain environments, than those that are not. As plasticity is said to increase with environmental variability [23], we investigate this by evolving agents that 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' similarly to [12], where agents must learn multiple stages of a task to achieve a goal. Further, we explore the effect that changing the context in which an agent exists has on evolution, by changing the environment from single- to multi-agent. We hypothesise that behavioural plasticity will help agents to achieve their tasks in these environments, by facilitating immediate behavioural changes in response to varying environmental contexts or conditions.

3 TESTBED AND AGENT DESIGN

3.1 The River Crossing Dilemma Testbed

The River Crossing Dilemma (RCD) testbed was introduced by Barnes et al. [5], to explore how agents evolve to achieve individual goals in shared worlds; this extends the original River Crossing Task proposed by Robinson et al. [33]. Agents must learn what their goal is and how to achieve it with no prior knowledge of the task or environment. The RCD is a 19×19 grid-world, with a two-cell deep river of Water. Each river bank has four Stones; all empty cells are Grass (Figure 1). An agent's goal is to collect its allocated Resources from either side of the river, which gives a highly positive fitness. Conversely, agents drown and receive a highly negative fitness when stepping into the river. The task is multi-stage [12], as agents must evolve to perform the appropriate behaviours in different contexts to achieve their goal: they must build a bridge to cross the river, and avoid drowning. Two Stones must be placed in the same Water cell to successfully build a bridge. Time is measured in 'timesteps'; an agent can move one cell per timestep. For experiments with two agents, the agent starting in the top left of the environment moves first, then the agent starting in the bottom right.

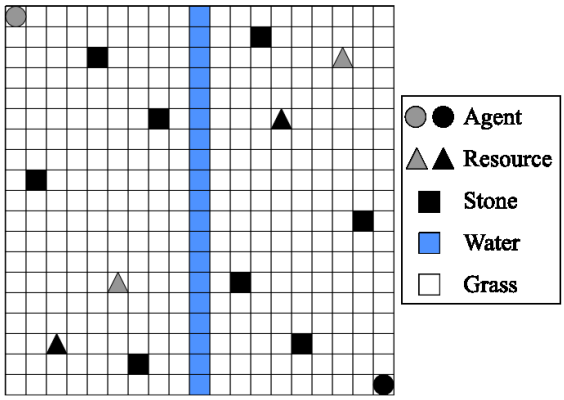


Fig. 1. The River Crossing Dilemma testbed, proposed by Barnes et al. [5]. The grey agent (top left) is allocated the two Resources in grey, and the black agent (bottom right) is allocated the two Resources in black; agents cannot interact with Resources not allocated to them. Both agents can interact with all other objects. For single-agent environments, the black agent is removed.

3.2 The Protected River Crossing Dilemma

We introduce the Protected River Crossing Dilemma (PRCD) – an adaptation of the RCD [5] specifically used to explore how agents evolve to solve single-stage tasks; like with the RCD, the PRCD is a Java implementation and is the same as Figure 1. However, the river acts as an impassable and non-lethal obstacle; agents cannot fall into it mistakenly. This simple change means that agents do not need to learn the different contexts in which they can interact with the river: it is not safe unless carrying a Stone. 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.

3.3 Gamification of the RCD and PRCD

The RCD and PRCD are gamified, such that agents incur an increasing, personal cost for each Stone placed in the river; a bridge is successfully built with two Stones, since the river is two cells deep. This cost introduces a social

dilemma in multi-agent environments, specifically a Snowdrift Game [30]; this means that agents may: complete their task individually and endure the full cost of bridge-building; cooperate to share the cost; or exploit other agents by waiting for them to build a bridge, to avoid a cost at all. In addition to creating a social dilemma, this increasing cost for placing Stones also deters agents from learning to simply place Stones in the river, encouraging them to achieve their goal with the least effort. We use the term gamification in a slightly broader way than is typically considered in, for example, the gamification of uninteresting tasks for people. In this paper, we simply mean use it to refer to the addition of game elements to a task. This gamification means that there is less incentive for agents to cooperate due to the cost of bridge-building, but defection can lead to failure if the agent isn't able to achieve its goal. The fitness, or payoff, for agent i is calculated with Equation 1:

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

where r_i is the number of Resources collected by agent i, $N = 2$ and is the number of Resources that an agent must collect in total to achieve its goal (each Resource is allocated to a specific agent to collect), $C = 0.1$ and is the cost of placing a Stone in the river, s_i is the number of Stones placed in the river by agent i, and $f = 1$ if agent i falls in the river, or 0 otherwise. An agent's fitness is calculated based on its *own* behaviour. Commonly observed fitnesses are presented in a payoff matrix in Table 1. Achieving the goal alone gives a fitness of 0.7, which increases to 0.9 if the cost of bridge-building is shared, or 1.0 if an agent exploits another; anything below 0.7 indicates the goal is not achieved.

3.4 Agent Design

Agents in both the RCD and the PRCD use a two-tiered neural network architecture, adapted from Barnes et al. [5] and inspired by Robinson et al. [33]. The first tier is the deliberative network, which generates high-level sub-goals based on the current inputs, corresponding to the agent's current state. This network is responsible for decision-making; depending on the inputs and weights of the network, the outputs indicate what the agent's current sub-goals are (whether it is attracted to, neutral towards or repulsed from certain objects). The weights of the network (as well as the type of each neuron) represent the agent's genes, 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). The 'partial bridge' input informs agents anywhere in the environment that a Stone has been placed somewhere in the river; this helps navigation efforts by indicating that some parts of the river are 'shallower' than others, and only require one more Stone to build a bridge. This feed-forward network has six input neurons, three hidden layers with eight, six and

Table 1. Payoff matrix using Equation 1 to show the fitness achieved by agent x in the RCD [5] and PRCD testbeds, assuming that agent x has retrieved both Resource objects and another agent y exists in the environment. S_x and S_y are the number of Stones placed by each Agent. $S_y = 0$ also shows the fitnesses that agent x could achieve if it exists in an environment alone.

	$S_y = 0$	$S_y = 1$	$S_y = 2$
$S_x = 0$	0.0	0.0	1.0
$S_x = 1$	-0.1	0.9	0.9
$S_x = 2$	0.7	0.7	0.7

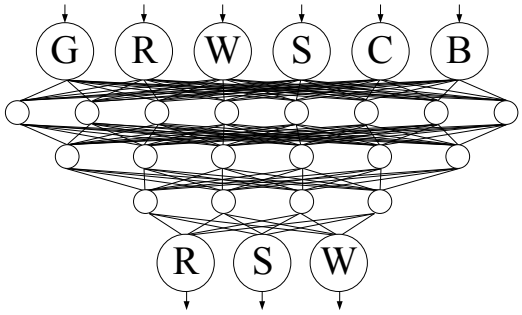


Fig. 2. The deliberative neural network has three hidden layers, and generates high-level sub-goals based on the current state. Inputs are 1 or 0, corresponding to the agent's current state: Grass (G), Resource (R), Water (W), Stone (S), Carrying Status (C), if a Bridge partially exists (B). Outputs are 1 for attraction, 0 for neutral or −1 for avoidance for each sub-goal: Resource (R), Stone (S), Water (W).

four neurons respectively, and an output layer of three neurons (Figure 2); each neuron in each layer of neurons in the network is connected to each of the neurons in the next layer. Resources, Stones and Water will be attractive if the output is 1, avoided if −1, or neutral if 0. Snell-Rood [35] posits that activational behavioural plasticity – the focus of this work – increases with brain size, in terms of the number of neurons; Herczeg et al. [19] observe this effect in guppies, where brain size can indicate the degree of plasticity and an individual's ability to adapt to novel environments. Increasing the number of neurons in the deliberative network in this study compared to prior work [5] is intended to increase the degree of plastic behaviour compared to a smaller network.

The second tier is the reactive network, with the same dimensions as the environment – in this case, 19×19; each neuron is connected to the surrounding eight. This reactive network uses the shunting equation (Equation 2, [5, 33, 44, 45]) to create dynamic activity landscapes based on the current sub-goals; the activity for each neuron, and thus the overall activity landscape, is calculated with this equation at each timestep, meaning agents can react immediately when their goals change. Agents can therefore hill-climb towards the goals generated in the previous tier by moving to the cell in its Moore neighbourhood (the surrounding eight cells) with the highest activity. Note that Equation 2 is used exclusively in the reactive network, not the deliberative network. Agents must make one move per timestep and cannot remain stationary. Agents also cannot move into a cell occupied by another agent. 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 adjacent cell 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 network (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 [33].

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

3.5 Operationalising Activity-Gating Neuromodulation

Modulated agents can immediately change their phenotype/behaviour by regulating and temporarily suppressing activity *within* the deliberative network (Figure 2) – without permanently changing network weights. Figure 3 shows an example of this activity-gating modulation. Neurons in the deliberative network may evolve to be non-modulatory or modulatory; they both propagate activity in the same, standard way – except when the incoming signal (sum of inputs) to a modulatory neuron is negative. In this case, the neuron will regulate activity (and ultimately behaviour) by outputting a signal of 0 along each of its outgoing connections. These connections are thus effectively 'turned-off', or gated, as the signal is blocked locally; note that the weights themselves are not changed. 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 network, is therefore able to express multiple phenotypes – without changing, or potentially destroying, the knowledge encoded in the weights. In other words, a modulatory agent can temporarily change 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.

3.6 Evolutionary Algorithm

All experiments are conducted using the PRCD and RCD testbeds with the following common parameters, inspired by [5]. 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 achieve their goal. As a steady state genetic algorithm is used, only one agent is replaced in the population at each timestep; a tournament of three randomly selected agents allows more areas of the solution space to be explored, whereas evaluating the whole population would restrict the search. 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 network), this offspring has a probability of $P_{one} = 0.95$ to inherit the chromosome from a random parent,

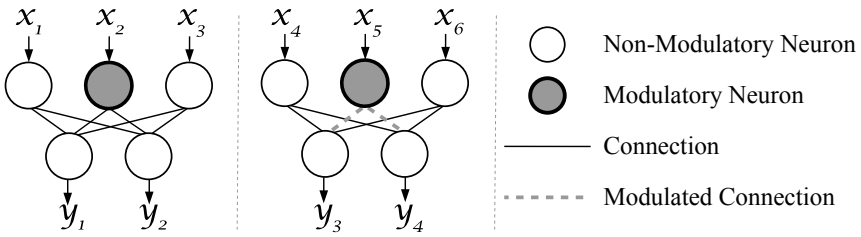


Fig. 3. Modulatory neurons in the deliberative network propagate activity the same as non-modulatory neurons when the sum of the neuron's inputs is ≥ 0; here, if the input signal to x_2 is positive, the outgoing activity signals of x_2 propagate through the connections to the next layer of neurons as usual (y_1 and y_2). If, however, the input signal to x_5 is negative, the modulatory neuron regulates the outgoing activity; specifically, 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 permanently modifying the network weights. Modulatory neurons only affect their own outgoing connections, so the connections from x_4 and x_6 to y_3 and y_4 are unaffected by x_5.

otherwise single-point crossover is used. Each connection weight w in the offspring's deliberative network is then mutated by a random value from a Gaussian distribution with $\mu = w$ and $\sigma = 0.01$.

For modulatory agents, the hidden neurons in the deliberative 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 deliberative network of each agent is initialised with non-modulatory neurons, then evolved with neuroevolution like the weights of the network. 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 hidden neuron 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 [13]. Modulatory neurons regulate activity as outlined in Section 3.5. Non-modulatory agents have a static network of non-modulatory neurons that do not evolve.

4 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; the experimental study is designed to explore how the ability to rapidly and reversibly change phenotypic behaviour helps agents to solve tasks in varying environmental conditions. All experiments are repeated 100 times, both with and without neuromodulation, and evolve agents for 500,000 generations from a randomly-initialised state unless otherwise specified.

The experiments in Section 5.1 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 later experiments. Variability increases if there is more than one agent in the environment, since the actions of each agent can change the environment unpredictably.

The second set of experiments introduces another agent into the single-stage task PRCD; this creates a social dilemma, so 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 is therefore harder to evolve in. These experiments evolve two separate, randomly-initialised populations of agents that start on opposite corners of the environment. In multi-agent environments, only the evolution and goal-achievement of the agent that begins in the top-left corner is analysed; this makes the results from single- and multi-agent environments comparable. The other agent still evolves as described in Section 3.6, however its evolution is not analysed.

The third set of experiments investigates how agents that exist alone evolve to solve a multi-stage task in 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 and the challenge of 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 [43].

Table 2. 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. Agents evolve alone, together, or with continued evolution (CE). 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 (S/M)	Fitness (% of Agents)				
		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

5 RESULTS

5.1 Learning Single-Stage Tasks When Alone

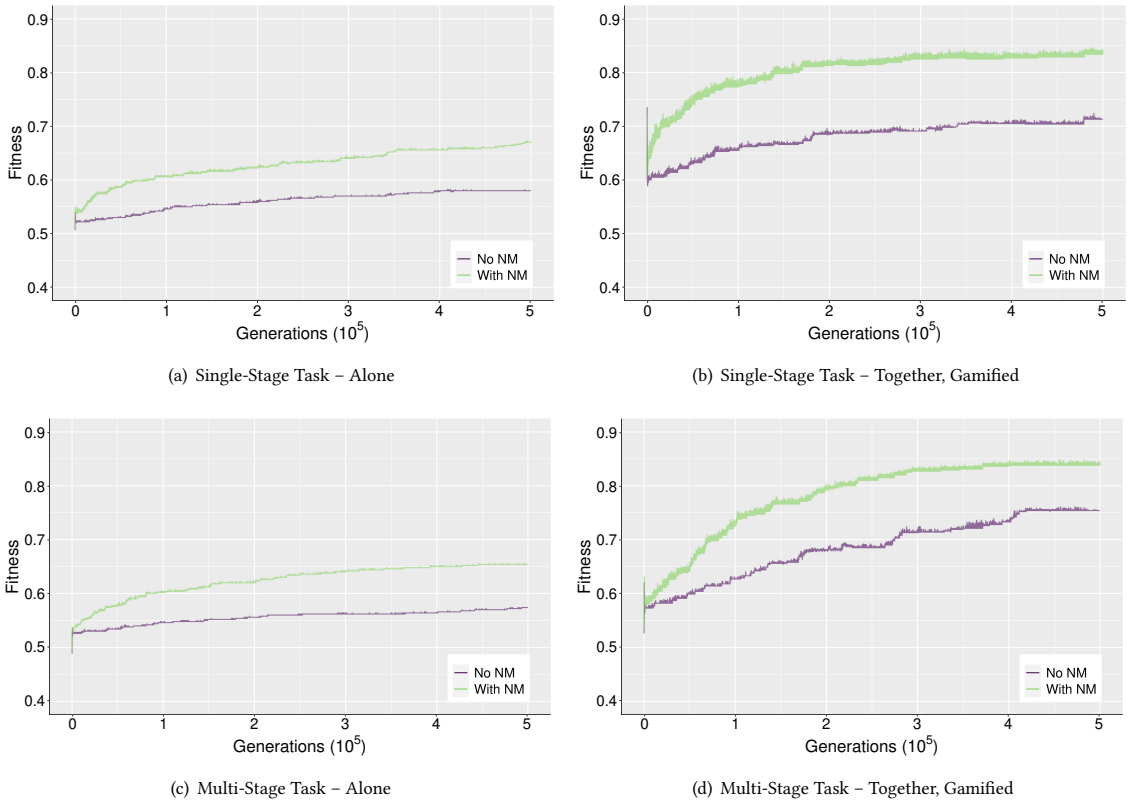
We start by investigating how agents evolve to solve the simplest task in the least variable environment in the study – the single-stage task in the Protected River Crossing Dilemma (PRCD) – and the role that neuromodulation plays.

Figure 4(a) shows the mean best-in-population fitness of agents evolving alone in the PRCD, both with and without neuromodulation. The benefit of neuromodulation is seen at the start of, and is sustained throughout, evolution. 85% of modulatory agents were able to achieve their goal at the end of evolution, compared to only 40% of non-modulatory agents (Table 2).

5.2 Learning Single-Stage Tasks When Together

The single-stage task in the PRCD becomes gamified when there are two agents; the actions of the other agent are unpredictable, meaning variability also increases. 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.

Figure 4(b) shows the mean best-in-population fitness of agents evolving together in a shared PRCD environment. Similarly to when agents evolve alone (Figure 4(a)), neuromodulation is beneficial from the start. Modulatory agents achieve higher fitnesses more often than their non-modulatory counterparts, and by the end of evolution, 97% of modulatory agents achieve their goal compared to 61% of non-modulatory agents (Table 2). The effect of neuromodulation is more prominent when agents evolve together compared to when they evolve alone, as agents can achieve a higher fitness. This finding is not interesting in itself, however the fact that fewer agents evolve to achieve their goals individually in shared environments (Table 2) demonstrates the impact that evolving in shared environments can have on goal-achievement. Relying on other agents to achieve goals can be detrimental if those agents change their behaviour or leave the environment. Further, the spike in fitness at the beginning of evolution is caused by both agents reacting to



(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- and multi-stage tasks take place in the PRCD and RCD respectively. 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).

and evolving based on the changes in the other's behaviour; once each agent's behaviour becomes more predictable, this spike drops. This is also observed in Figure 4(a).

5.3 Learning Multi-Stage Tasks When Alone

The multi-stage task in the RCD creates a more variable environment than in the single-stage task PRCD; 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, due to the bridge-building cost (Equation 1).

The mean best-in-population fitness increases over time as more agents evolve successful solutions; after 500,000 generations, 37% of agents achieved their goal without neuromodulation, compared to 77% with neuromodulation (Table 2). Figure 4(c) shows that the mean best-in-population fitness is higher 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.

5.4 Learning Multi-Stage Tasks When Together

The fitness function presented in Equation 1 evaluates each agent individually. In a shared environment, agents can still achieve their goal alone by building a bridge completely by themselves and enduring the associated cost; they can also exploit the other to avoid the cost, or cooperate to share the cost of bridge-building. The maximum fitness 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, agents perceive changes in environmental stimuli, and attempt to adapt their behaviour accordingly. The multi-stage task 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.

One thing to note about the difference between evolving in single- and multi-agent environments is that agents in multi-agent environments are affected by the actions of the other agent in one way or another; this is seen when agents solve single- and multi-stage tasks. Table 2 shows that fewer agents achieve their goals individually (by building a bridge on their own, to receive a fitness of 0.7) when evolving together, than when evolving alone; overall, more agents achieve their goals in shared environments because some exploit or cooperate with the other agent, but this may be detrimental in the long run if agents are unable to learn bridge-building behaviour themselves.

Figure 4(d) shows that modulatory agents evolve to achieve their goal more often, and in fewer generations, than non-modulatory agents. After 500,000 generations, 94% of modulatory agents achieve their goal, compared to only 68% of non-modulatory agents (Table 2). This shows that agents receive a benefit from expressing behavioural plasticity in response to changes in environmental stimuli caused by the actions of others.

5.5 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 utilise 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. 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 individual to shared, 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 mean best-in-population fitness for agents that continue to evolve together. The change in context from a single- to a multi-agent environment allows agents to immediately capitalise on the actions of others to achieve a higher fitness; Figure 6 shows the 5,000 generations either side of the context change at generation 500,000, which clearly shows a jump in fitness. This spike then falls slightly while agents adjust to the new change in context. Agents evolve in tandem and change their behaviour in response to the other agent's changes in behaviour; the spike then falls slightly as agents learn that other agents might not always be reliable, and thus evolve to achieve lower fitness by achieving goals alone. Neuromodulation is observed to help agents to adapt to their new, shared environment when

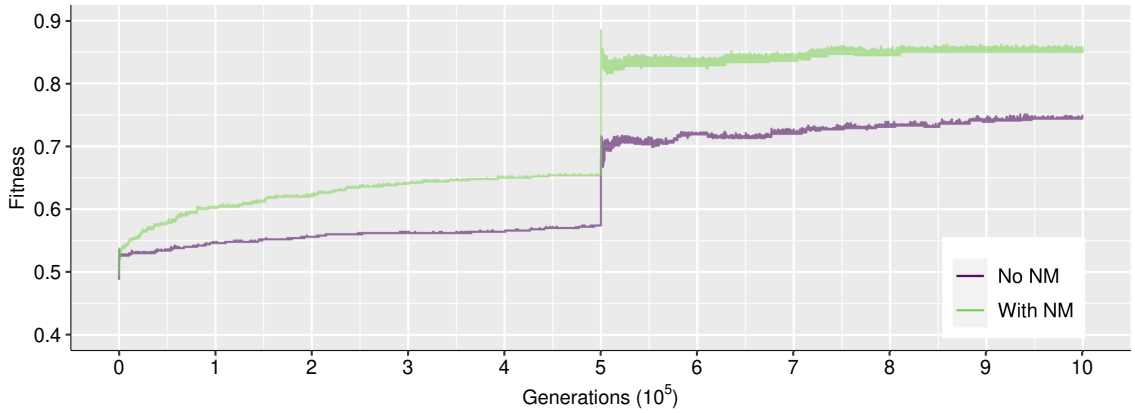


Fig. 5. The mean best-in-population fitness of agents that evolve alone for 500,000 generations, then continue to evolve together (Continued Evolution (CE)) with a partner for a further 500,000 generations, with and without neuromodulation (NM). A fitness of 0.7 or above indicates the goal is achieved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

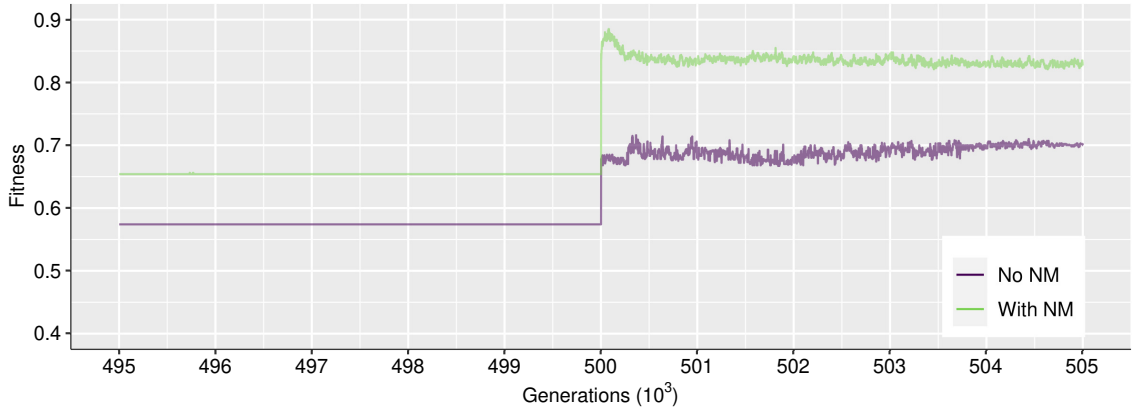


Fig. 6. The mean best-in-population fitness of agents evolving with continued evolution (CE), with and without neuromodulation (NM) – 5,000 generations before and after the change from a single- to multi-agent environment. A fitness of 0.7 or above indicates the goal is achieved (Equation 1).

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 (Table 2).

6 ANALYSING THE EFFECT OF BEHAVIOURAL PLASTICITY AND ENVIRONMENTAL VARIABILITY ON AGENT EVOLUTION

In these experiments, activity-gating neuromodulation increases both the likelihood and the speed that agents evolve successful solutions – both when they exist alone, and when they exist together (Figure 4). This section analyses agent evolution further, and aims to ascertain whether behavioural plasticity affects the fitness agents receive both at the end

Table 3. Statistical moments and median (to 3 S.F.) of the best-in-population fitness after 500,000 generations of evolving alone, together, and with continued evolution (CE). The highest mean and median, and lowest amount of skewness, kurtosis and variance for each experiment with and without neuromodulation are in **bold**.

Exp	Task	NM	Mean	Median	Skewness	Kurtosis	Variance
Alone	S	N	0.58	0.5	**0.408**	**1.17**	0.00970
		Y	**0.67**	**0.7**	-1.96	4.84	**0.00515**
	M	N	0.574	0.5	**0.539**	**1.29**	0.00942
		Y	**0.654**	**0.7**	-1.28	2.647	**0.00716**
Together	S	N	0.713	0.7	0.345	1.548	0.0422
		Y	**0.836**	0.7	**-0.0856**	**1.424**	**0.0252**
	M	N	0.754	0.7	**0.0245**	**1.364**	0.0447
		Y	**0.838**	**0.85**	-0.282	1.60	**0.0286**
CE	M	N	0.744	0.7	0.195	1.61	0.0385
		Y	**0.852**	**0.95**	**-0.132**	**1.21**	**0.0233**

of and during evolution. Additionally, we explore whether neuromodulation affects the volatility that agents experience in terms of fluctuations in fitness during evolution.

6.1 Analysing the Fitness

The statistical moments and median for the best-in-population fitness of each experiment were calculated at the end of evolution (Table 3). This analysis shows that modulatory agents have the same or higher mean and median fitness across all experiments. Combined with the results presented in Table 2, modulatory agents not only have a higher mean and median fitness, but they achieve their goal more often than non-modulatory agents; this is observed both in single- and multi-stage tasks and single- and multi-agent environments. The variance in the best-in-population fitness after evolution is also lower in modulatory agents, which further illustrates the benefits of behavioural plasticity.

The distribution of fitnesses after evolution for modulatory agents is negatively skewed; the amount of skewness tends to decrease from highly skewed to more symmetrical as environmental variability increases. This is supported by the median fitness tending to be higher than the mean fitness for modulatory agents, meaning that agents would likely achieve a higher-than-average fitness. The opposite is observed in non-modulatory agents, as the fitness distribution is positively skewed; as with modulatory agents, the amount of skew tends to decrease as environmental variability increases. In each experiment, the mean fitness for non-modulatory agents is higher than the median; this indicates positive skewness, and that agents would be likely to achieve a fitness lower than the average. A contributing factor to this is that non-modulatory agents are less likely to evolve a goal-achieving fitness at the end of evolution than modulatory agents, thus skewing the distribution of fitnesses to the left.

The amount of kurtosis in the fitness distribution tends to increase in non-modulatory agents as environmental variability increases, but decrease in modulatory agents; this suggests that more outliers can be expected in non-modulatory agents as environmental variability increases, and the opposite in modulatory agents. Saying this, all fitness distributions for each experiment are platykurtic (where excess kurtosis is negative ($kurt_{excess} = kurt - 3$), or $kurt < 3$), meaning that outliers and extreme values are not common overall.

To analyse the effect that activity-gating neuromodulation has on evolution further, statistical tests were performed to compare the best-in-population fitnesses of modulatory and non-modulatory agents in each experiment. Firstly, a

Table 4. Wilcoxon Signed Rank Statistical Tests comparing the fitness and volatility metrics (Section 6.3) of the best-in-population non-modulatory (m_n) and modulatory (m_m) agents in 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 indicated with an asterisk (*).

Metric	Exp	Task (S/M)	Statistical Test Alternative Hypothesis		
			$m_n \neq m_m$	$m_n < m_m$	$m_n > m_m$
Fitness	Alone	S	2.588×10^{-9} *	1.294×10^{-9} *	1
	Together	S	2.362×10^{-4} *	1.181×10^{-4} *	0.9999
	Alone	M	2.994×10^{-8} *	1.497×10^{-8} *	1
	Together	M	1.594×10^{-2} *	7.97×10^{-3} *	0.9922
	CE	M	2.593×10^{-4} *	1.296×10^{-4} *	0.9999
SDoT	Alone	S	1.297×10^{-4} *	6.484×10^{-5} *	0.9999
	Together	S	4.535×10^{-3} *	2.267×10^{-3} *	0.9978
	Alone	M	2.052×10^{-7} *	1.026×10^{-7} *	1
	Together	M	6.919×10^{-2}	3.46×10^{-2} *	0.9657
	CE	M	2.315×10^{-2} *	1.157×10^{-2} *	0.9885
CACoT	Alone	S	5.881×10^{-5} *	2.941×10^{-5} *	1
	Together	S	3.459×10^{-10} *	1.73×10^{-10} *	1
	Alone	M	3.982×10^{-4} *	1.991×10^{-4} *	0.9998
	Together	M	1.371×10^{-6} *	6.856×10^{-7} *	1
	CE	M	1.213×10^{-5} *	6.064×10^{-6} *	1
CCoT	Alone	S	5.699×10^{-5} *	2.849×10^{-5} *	1
	Together	S	3.52×10^{-8} *	1.76×10^{-8} *	1
	Alone	M	4.028×10^{-4} *	2.014×10^{-4} *	0.9998
	Together	M	3.152×10^{-5} *	1.576×10^{-5} *	1
	CE	M	1.982×10^{-6} *	$9,908 \times 10^{-7}$ *	1

Shapiro-Wilk test for normality is described by Yap and Sim [47] as being powerful for a range of distributions that are skewed, symmetric, and those with high or low kurtosis. As such, it is appropriate to test the distributions described in Table 3. Each distribution was found to be non-normal ($p < 0.05$).

As the distributions are non-normal, Wilcoxon Signed Rank statistical tests were then conducted to analyse the effects of behavioural plasticity on fitness and evolution. This non-parametric test compares 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 4. The two-tailed tests show that there is a significant difference in median fitness between non-modulatory and modulatory 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 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 a positive effect on the expected fitness of agents, in all areas of the study.

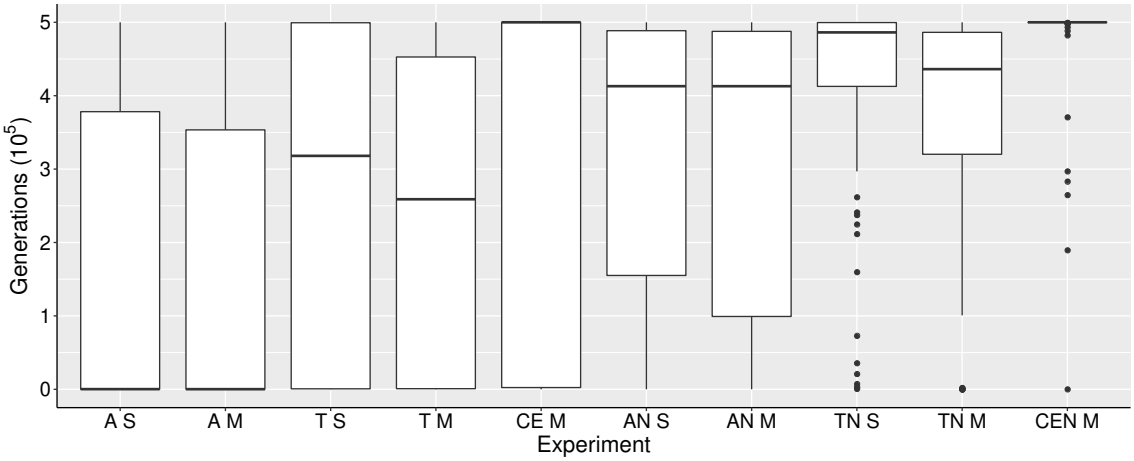


Fig. 7. Box plot depicting the number of generations agents receive a goal-achieving fitness (≥ 0.7) during 500,000 generations of evolution in each experiment: alone (A), together (T), with continued evolution (CE, excluding the initial period of evolving alone), with neuromodulation (N), in a single- (S) or multi- (M) stage task.

6.2 Analysing Goal-Achievement over Evolution

Thus far, the fitness that agents receive at the end of evolution has been assessed; modulatory agents are observed to have a higher mean fitness than non-modulatory agents, and achieve their goals more often. However, while it is desirable to evolve agents that can receive a goal-achieving fitness at the end of evolution, another benefit would be for agents to consistently achieve their goals throughout evolution as well.

Figure 7 shows a box plot of the number of generations that agents receive a goal-achieving fitness (≥ 0.7) during evolution. In each experiment, the first, second and third quartiles of goal-achieving generations is the same or higher in modulatory agents than in non-modulatory agents; this shows that modulatory agents achieve their goal for more generations overall than their non-modulatory counterparts. Not only is the data more heavily skewed to the left in modulatory agents, but the spread of values is generally smaller than in non-modulatory agents; this indicates that modulatory agents are more predictable, and are likely to spend more generations receiving a goal-achieving fitness than other agents. Modulatory agents thus spend more of their lifetime able to achieve their goals than agents not capable of behavioural plasticity.

To evidence this claim further, Wilcoxon Signed Rank statistical tests were conducted to compare the number of successful generations between non-modulatory and modulatory agents, where a 'successful' generation is one that an agent receives a goal-achieving fitness of ≥ 0.7 (Table 5). In line with Section 6.1, a Shapiro-Wilk normality test first indicated each distribution was non-normal ($p < 0.05$). In each experiment, the two-tailed Wilcoxon Signed Rank test shows that there is a significant difference in the median number of successful generations between non-modulatory and modulatory agents; as $p < 0.05$ in each test, the null hypothesis that the medians are equal can be rejected. Further, the one-tailed tests show that there is a significant directional difference between the two medians, where the median number of successful generations in non-modulatory agents is lower than in modulatory agents. The analysis thus far therefore shows that behavioural plasticity can help agents to not only be more likely to receive a higher fitness and achieve their goals after evolution, but it can also help them to be more successful throughout evolution as well.

Table 5. Wilcoxon Signed Rank Statistical Tests comparing the number of generations that the best-in-population non-modulatory (m_n) and modulatory (m_m) agents receive a goal-achieving fitness (≥ 0.7) in each experiment: evolving 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 indicated with an asterisk (*).

Exp	Task (S/M)	Statistical Test Alternative Hypothesis		
		$m_n \neq m_m$	$m_n < m_m$	$m_n > m_m$
Alone	S	2.128×10^{-7} *	1.064×10^{-7} *	1
Together	S	1.846×10^{-7} *	9.23×10^{-8} *	1
Alone	M	3.031×10^{-7} *	1.515×10^{-7} *	1
Together	M	2.177×10^{-6} *	1.088×10^{-6} *	1
CE	M	2.484×10^{-8} *	1.242×10^{-8} *	1

6.3 Analysing the Effect of Behavioural Plasticity on Evolutionary Volatility

Behavioural plasticity arising through neuromodulation has a positive impact on the fitness agents achieve after evolution, and the ability of agents to achieve their goals. In this section, we explore how behavioural plasticity affects the evolution and thus the evolved fitness of agents.

Barnes et al. [6] proposed three metrics to analyse the volatility of agent evolution, by capturing the variability and dispersion of values over time. These metrics can therefore be used to describe the evolutionary process of agents, and whether the received fitness is prone to change frequently during evolution.

The Standard Deviation over Time (SDoT) metric is inspired by a common metric used in volatility forecasting in finance, capturing the dispersion and variability of values over time by calculating the sample standard deviation over a defined time period. A high SDoT indicates that agents have highly volatile evolution, meaning that the fitness has a high variability and dispersion of values over time.

The Cumulative Absolute Change over Time (CACoT) metric is used to analyse how much an agent's fitness fluctuates over time, by capturing the magnitude of fitness changes during evolution; an agent whose fitness fluctuates by large amounts would therefore have a high CACoT. This is calculated by totalling the absolute change in fitness between each generation.

Complementary to the previous metric, the Count of Change over Time (CCoT) metric captures how often an agent's fitness changes from one generation to the next during evolution – without capturing the magnitude of the changes; a high CCoT indicates that the fitness changes often.

For all 100 runs of each experiment, a value for each of the three metrics was calculated using the best-in-population fitness at each generation across 500,000 generations of evolution, or all 1,000,000 generations for agents evolving with Continued Evolution. Statistical moments and medians are presented for each metric in Tables 6, 7 and 8.

In all experiments, non-modulatory agents have a lower mean and median SDoT, CACoT and CCoT (Tables 6, 7 and 8 respectively) than their modulatory counterparts, indicating that evolution is more volatile for modulatory agents and that the received fitness tends to fluctuate often. The increase in volatility when agents share an environment can be observed in Figures 4 and 5, since the line graphs appear 'thicker' than when agents evolve alone; this is because the fitness fluctuating often. This volatility would partly be caused by agents reacting to the other agent's behaviour, which may potentially be different to the previous generation; it could also be due to the mutations that occur at each generation, which would make the effect of neuromodulation stronger or weaker depending on the strength of the

Table 6. Statistical moments and median (to 3 S.F.) of the SDoT volatility metric for the best-in-population agents after 500,000 generations of evolving alone, together, and with continued evolution (CE). The lowest values for each experiment with and without neuromodulation are in **bold**.

Exp	Task	NM	Mean	Median	Skewness	Kurtosis	Variance
Alone	S	N	**0.0251**	**0.00306**	1.10	2.40	0.00134
		Y	0.0443	0.0374	**0.254**	**1.56**	**0.00131**
	M	N	**0.0188**	**0.00285**	1.55	3.74	**0.00101**
		Y	0.0419	0.0382	**0.292**	**1.49**	0.00139
Together	S	N	**0.0489**	**0.0157**	1.65	4.94	0.00371
		Y	0.0665	0.0537	**1.19**	**3.89**	**0.00315**
	M	N	**0.0802**	**0.0310**	0.763	**2.07**	0.00735
		Y	0.0981	0.0804	**0.636**	2.16	**0.00594**
CE	M	N	**0.102**	**0.0616**	0.507	**1.56**	0.0105
		Y	0.130	0.133	**0.159**	1.59	**0.00690**

Table 7. Statistical moments and median (to 3 S.F.) of the CACoT volatility metric for the best-in-population agents after 500,000 generations of evolving alone, together, and with continued evolution (CE). The lowest values for each experiment with and without neuromodulation are in **bold**.

Exp	Task	NM	Mean	Median	Skewness	Kurtosis	Variance
Alone	S	N	**4.28**	**1.70**	5.95	45.3	**72.2**
		Y	8.13	4.70	**5.70**	**44.4**	174
	M	N	**2.79**	**1.1**	3.25	14.7	**18.0**
		Y	4.37	2.30	**2.57**	**9.59**	29.5
Together	S	N	**41.4**	**22.7**	**1.55**	**5.18**	**1730**
		Y	206	116	2.69	13.5	69600
	M	N	**40.8**	**13.3**	5.15	37.2	**6560**
		Y	97.1	46.1	**3.38**	**19.0**	16400
CE	M	N	**60.6**	**8.65**	5.97	45.5	**25800**
		Y	231	47.0	**2.68**	**11.8**	149000

mutated connections in the deliberative network. Further, agents have a lower mean and median CACoT and CCoT when evolving to solve a multi-stage task than a single-stage task, both with and without neuromodulation. The results therefore suggest that the best-in-population fitness fluctuates less and by lower amounts during evolution when agents solve a multi-stage task compared to a single-stage task. The exception is that the mean CCoT of non-modulatory agents evolving together is higher for the multi-stage task than the single-stage task. A similar trend can be seen in Table 2, as more agents solve the single-stage task than the multi-stage version – except when non-modulatory agents evolve together; this would result in more fluctuations in fitness during evolution, and a higher CCoT.

Non-modulatory agents have lower variability in CACoT and CCoT, however modulatory agents generally have a lower variability in SDoT. These findings, combined with a lower mean and median in each metric, indicate that non-modulatory agents have fewer and more predictable fluctuations in fitness with less magnitude, and a higher and less predictable SDoT than in modulatory agents. Additionally, the mean, median and variance for each metric tend to

Table 8. Statistical moments and median (to 3 S.F.) of the CCoT volatility metric for the best-in-population agents after 500,000 generations of evolving alone, together, and with continued evolution (CE). The lowest values for each experiment with and without neuromodulation are in **bold**.

Experiment	Task	NM	Mean	Median	Skewness	Kurtosis	Variance
Alone	S	N	**19.9**	**7.00**	5.94	45.3	**1810**
		Y	39.1	22.0	**5.70**	**44.4**	4340
	M	N	**12.5**	**4.00**	3.25	14.7	**449**
		Y	20.3	10.0	**2.58**	**9.61**	738
Together	S	N	**155**	**75.5**	**2.52**	**11.6**	**39600**
		Y	854	356	2.85	14.1	1690000
	M	N	**174**	**35.0**	**3.38**	**15.2**	**129000**
		Y	373	134	4.00	24.2	360000
CE	M	N	**321**	**32.5**	8.05	73.3	**1530000**
		Y	1130	228	**2.58**	**10.9**	3470000

increase as environmental variability increases; the results therefore suggest that agents will experience more volatility as environmental variability increases, where volatility is likely to: be lowest in agents that evolve alone; increase when agents evolve together; and be highest when agents evolve with continued evolution.

Each metric for each experiment has positive skewness, showing that the data is right-skewed; this is supported by the median being lower than the mean (except for a marginally higher median than mean for the SDoT of modulatory agents evolving with continued evolution in a multi-stage task (Table 6)). The CACoT and CCoT distributions for each experiment are highly skewed, whereas the SDoT is generally less skewed. Positive skewness indicates that agents would likely have a lower SDoT, CACoT and CCoT than the average, as the distribution is skewed by higher values; agents would therefore be expected to have a lower CACoT and CCoT than the observed mean and median. Further, the skewness and kurtosis of each metric is generally lower in modulatory agents than in non-modulatory agents; the values for each metric would be less likely to be extreme and more likely to be symmetrical around the mean, with outlier values being less likely in modulatory agents than non-modulatory agents.

Further to the analysis of fitness in Section 6.1, a Shapiro-Wilk test was conducted to detect normality in the SDoT, CACoT and CCoT distributions for each experiment; $p < 0.05$ for each test, indicating non-normality. Wilcox Signed Rank statistical tests (one two-tailed ($m_n \neq m_m$), and two one-tailed tests ($m_n < m_m$, $m_n > m_m$)) were then performed; the results are presented in Table 4. The two-tailed tests show that for each experiment, there is a significant difference between the metric for non-modulatory and modulatory agents (except for the SDoT of agents evolving together to solve a multi-stage task), as $p < 0.05$ in each test. Further, the results of the one-tailed tests with the alternative hypothesis $m_n < m_m$ show that for each experiment, there is a significant directional difference in the medians of the two distributions, where the metric for non-modulatory (m_n) agents is significantly lower than modulatory agents (m_m); each p-value is below 0.05, thus the null-hypothesis that there is no directional difference in medians can be rejected. The final one-tailed tests with the alternative hypothesis $m_n > m_m$ show no significant difference.

Overall, this analysis shows that modulatory agents experience more evolutionary volatility, which also tends to increase with environmental variability; as the environment gets more unpredictable and uncertain due to the unknowable actions of others, fitness tends to fluctuate more. There does however seem to be a trade-off between

Table 9. The most common number of modulatory neurons evolved in each of the three layers of the deliberative networks (L1, L2, L3), and in total (LT). Results are presented for agents evolving to solve a single- (S) or multi- (M) stage task, and those that achieve (Y) their goal (G) and those that do not (N). The frequency that the configuration occurs is shown, as well as the total number of agents overall. A dash (-) indicates that no configuration occurred more than once.

Experiment	Task	G	L1	L2	L3	LT	Freq	Total
Alone	S	Y	4	3	3	10	6	85
		N	3	3	3	9	2	15
	M	Y	3	2	2	7	5	77
		N	3	2	3	8	2	23
Together	S	Y	4	4	2	10	6	97
		N	-	-	-	-	-	3
	M	Y	4	3	3	10	6	94
		N	-	-	-	-	-	6
CE	M	Y	5	4	2	11	8	99
		N	-	-	-	-	-	1

fitness and volatility; despite this higher level of evolutionary volatility, modulatory agents are observed to have a higher mean fitness than non-modulatory agents (Table 4), and achieve their goals more often.

6.4 Analysing the Modulatory Neurons in the Neural Networks

To understand the effect of behavioural plasticity via neuromodulation further, the arrangement of modulatory neurons that evolve in the agents were examined to see whether any patterns emerge. For each of the 100 runs of each experiment, the deliberative network for the single best-in-population agent after evolution was recorded for comparison.

Table 9 presents the most common configuration of modulatory neurons evolved in the deliberative networks in each experiment, broken down into agents that do and do not achieve the goal. It is worth noting that the frequency of these common configurations is low in comparison to the total number of agents that have and have not achieved their goal (e.g. six agents had a common configuration out of 85 that achieved their goal when evolving alone to solve a single-stage task). As such, no configuration leads agents to achieve their goal or not.

It is therefore apparent that agents can achieve their goal in many different ways, with different numbers of modulatory neurons in each layer and in different arrangements. It is not clear whether all modulatory neurons in these configurations are used or beneficial – some may be redundant if the surrounding weights are near zero values. Saying this, no agent was observed to evolve a neural network with either zero modulatory neurons or the maximum out of a possible 18 – each agent evolved a deliberative neural network with at least three modulatory neurons. This suggests that there is no obvious link between the number or configuration of modulatory neurons and either the success of an agent, the behaviours that the agent switches between, the stimuli that affects when modulation occurs, the type of environment it evolves in, or the task in which it has to solve. Because modulatory neurons can regulate neural network activity locally, this can potentially make goal-achieving behaviours (such as moving towards Water when a Stone is being carried, potentially by-passing the need to learn the negative association with the river) become accessible early on in evolution – without the agent needing to encode that exact knowledge directly in the network. This could be an explanation of why neuromodulation increases the mean best-in-population faster than in non-modulatory

agents in Figures 4 and 5. Further, agents did not converge to one single 'successful' or 'unsuccessful' configuration of modulatory neurons – modulatory neurons can be arranged in a number of different ways to have a positive effect on agent evolution and fitness.

7 DISCUSSION AND IMPLICATIONS

Evolving to solve tasks in dynamic and uncertain environments can be difficult for neural networks, as learning or altering behaviours in response to environmental changes means that knowledge encoded in the network will be changed; if this happens, goal-achieving behaviour may be lost, and fitness may degrade as a consequence. Barnes et al. [5] demonstrate the implications that pursuing individual goals in a shared environment can have when evolving neural networks as agent controllers. When two agents – each unaware of the other – act within a shared environment, these actions can interfere with the other's learning and ability to achieve their own goals. Unintended interactions can have an unexpected impact on how well suited an agent or system is for the environment it is located in. These issues are becoming evermore important to consider when designing technical systems, as they are increasingly being composed of many components or sub-systems. As a result, it becomes increasingly likely that these systems will interact in unintended and unpredictable ways [17], which can lead to the state of the environment changing without warning through the actions of others. The ability for a system to behave appropriately regardless of unexpected interactions or changes in environmental states therefore becomes crucial. In nature, animals and humans exhibit behavioural plasticity when faced with unknown situations; this allows biological systems to temporarily perform different behaviours to those that have been learnt in an attempt to survive or overcome environmental changes. We have therefore investigated how plasticity could affect systems that evolve to solve tasks in uncertain environments.

In this study, we have abstracted this concept by exploring how simulated agents evolve in varying environments, where variability increases in terms of the complexity of the task, and whether another agent can affect the environment with its actions. In all cases, agents have no knowledge of others and therefore cannot intend to interact with others – nor can they learn that a perceived environmental change is caused by the actions of another agent. We show that behavioural plasticity in the form of evolving with neuromodulation has a positive effect on the fitness that agents receive throughout and at the end of evolution. However, what may not be so intuitive is that behavioural plasticity also increases the volatility of agent evolution. Whilst fitness is higher in modulatory agents than non-modulatory agents overall, the fitness over evolution fluctuates more. One might expect that dynamic behaviour would – in addition to improving the chance of success – actually decrease the volatility within the system, by counteracting any dynamics or volatility present within the environment. The three metrics used to measure the amount of evolutionary volatility (SDoT, CACoT, and CCoT, Tables 6, 7 and 8 respectively) show this is not the case, as the mean and median for each metric tends to increase as the variability in the environment increases. Even in the least variable environments in the study, plastic agents experience more evolutionary volatility than agents that are not.

The findings and analyses presented in this article show that consequences can exist for systems that are unable to act appropriately in environments that are prone to change, or those that are shared. In reality, as systems and the components they are composed of grow larger, the opportunities for unintended interactions with others and unexpected environmental changes also increase. Behavioural plasticity is one route to equipping systems with the ability to overcome environmental uncertainty, although we have shown that this leads to more evolutionary volatility. Higher volatility is the cost for an increase in fitness and ability to achieve goals; further, agents spend more of their lifetime or evolution able to achieve their goals than those that are not capable of behavioural plasticity.

8 CONCLUSION AND FUTURE WORK

Increasing environmental variability makes learning challenging for neural controllers, as encoded information must be overwritten to learn new things when environmental conditions change. The capacity to immediately and temporarily change behaviour based on environmental stimuli is said to promote adaptation in variable environments [35, 43]. We have thus investigated the effect that activity-gating neuromodulation has on an agent's ability to evolve to succeed in environments of increasing variability, by exploring how agents evolve to solve both single- and multi-stage tasks in single- and multi-agent environments. An important element of this study is that agents cannot learn about the actions or existence of other agents; in this way, they cannot intend to cooperate or exploit one another.

This study uses the River Crossing Dilemma testbed [5] to explore how agents evolve to solve multi-stage tasks; additionally, we propose a new adaptation called the Protected River Crossing Dilemma to observe how agents evolve to solve single-stage tasks. Our results demonstrate that activity-gating neuromodulation has a significant 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 change behaviour in novel environments or situations. We also show that neuromodulation helps agents to adapt to new contexts and environmental changes, and that they can achieve their goals in many ways; no single arrangement of modulatory neurons influences an agent's ability to achieve goals in any of the experiments conducted.

Using three metrics to analyse evolutionary volatility, we show that the fitness of modulatory agents fluctuates significantly more than non-modulatory agents; this higher volatility is a result of modulatory neurons regulating activity in the neural networks in response to changing environmental stimuli. Despite this volatility, modulatory agents are more likely to achieve their goals and more often during evolution, and receive a higher fitness than non-modulatory agents. Behavioural plasticity arising from neuromodulation therefore creates a trade-off, as a significantly higher fitness and chance of goal-achievement comes at the cost of higher evolutionary volatility. This may indeed be desirable for agents that exist in highly unpredictable and unknown environments, by equipping them with the ability to respond quickly and appropriately to environmental change in a way that preserves or even improves fitness or performance.

The most variable environment in this study evolved agents alone for an initial period of time, before continuing to evolve them with another agent; future studies will investigate the extent to which behavioural plasticity enables agents to maintain goal-achieving behaviours when the presence of other agents is unpredictable. A limitation of this study is that a maximum of two agents are observed in any environment; exploring how more agents interact and evolve together in the future would give further insight into the consequence of unintended interactions, and how agents may evolve to overcome these to achieve their goals. Additionally, a future line of research arising from this study surrounds whether an agent could retain goal-achieving behaviour when a partner enters or leaves the environment unpredictably; this would test the limits or benefits of neuromodulation further, and give additional understanding about the interactions and consequences of evolving systems in shared and dynamic environments.

As systems are increasingly located in unpredictable and variable environments, possessing the ability to behave appropriately in unseen scenarios is evermore important. This study demonstrates that activity-gating neuromodulation allows agents to temporarily change behaviour in response to environmental changes without affecting knowledge encoded in their neural networks. Whilst behavioural plasticity is shown to improve fitness and goal-achievement, we also demonstrate that a trade-off exists as agents experience more volatility during evolution as a result.

ACKNOWLEDGMENTS

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] Larry F. Abbott. 1990. Modulation of Function and Gated Learning in a Network Memory. *Proceedings of the National Academy of Sciences of the United States of America* 87, 23 (1990), 9241–9245. https://doi.org/10.1073/pnas.87.23.9241

[2] Larry F Abbott and Sacha B Nelson. 2000. Synaptic Plasticity: Taming the Beast. *Nature Neuroscience* 3, 11 (2000), 1178–1183.

[3] Derrik E. Asher, Andrew Zaldivar, Brian Barton, Alyssa A. Brewer, and Jeffrey L. Krichmar. 2012. Reciprocity and Retaliation in Social Games With Adaptive Agents. *IEEE Transactions on Autonomous Mental Development* 4, 3 (2012), 226–238. https://doi.org/10.1109/TAMD.2012.2202658

[4] Chloe M. Barnes, Anikó Ekárt, Kai Olav Ellefsen, Kyrre Glette, Peter R. Lewis, and Jim Tørresen. 2020. Coevolutionary Learning of Neuromodulated Controllers for Multi-Stage and Gamified Tasks. In *Proceedings of the IEEE 1st International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*. IEEE, 129–138. https://doi.org/10.1109/ACSOS49614.2020.00034

[5] Chloe M. Barnes, Anikó Ekárt, and Peter R. Lewis. 2019. Social Action in Socially Situated Agents. In *Proceedings of the IEEE 13th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*. IEEE, 97–106. https://doi.org/10.1109/SASO.2019.00021

[6] Chloe M. Barnes, Anikó Ekárt, and Peter R. Lewis. 2020. Beyond Goal-Rationality: Traditional Action Can Reduce Volatility in Socially Situated Agents. *Future Generation Computer Systems* 113 (2020), 579–596. https://doi.org/10.1016/j.future.2020.07.033

[7] Jean M Barnes and Benton J Underwood. 1959. Fate of First-List Associations in Transfer Theory. *Journal of experimental psychology* 58, 2 (1959), 97–105. https://doi.org/10.1037/h0047507

[8] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. 2020. Learning to Continually Learn. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*. IOS Press, 992–1001. https://doi.org/10.3233/FAIA200193

[9] James M. Borg, Alastair Channon, Charles Day, et al. 2011. Discovering and Maintaining Behaviours Inaccessible to Incremental Genetic Evolution Through Transcription Errors and Cultural Transmission. In *Advances in Artificial Life: Proceedings of the 11th European Conference on the Synthesis and Simulation of Living Systems (ECAL 2011)*. MIT Press, 101–108. https://doi.org/10.7551/978-978-0-262-29714-1-ch019

[10] John A. Bullinaria. 2007. Understanding the Emergence of Modularity in Neural Systems. *Cognitive Science* (2007). https://doi.org/10.1080/15326900701399939

[11] Anurag Reddy Daram, Dhireesha Kudithipudi, and Angel Yanguas-Gil. 2019. Task-Based Neuromodulation Architecture for Lifelong Learning. In *Proceedings of the 20th International Symposium on Quality Electronic Design (ISQED)*. 191–197. https://doi.org/10.1109/ISQED.2019.8697362

[12] Amir Dezfouli and Bernard W. Balleine. 2019. Learning the Structure of the World: The Adaptive Nature of State-Space and Action Representations in Multi-Stage Decision-Making. *PLOS Computational Biology* 15, 9 (09 2019), 1–22. https://doi.org/10.1371/journal.pcbi.1007334

[13] Kai Olav Ellefsen, Jean Baptiste Mouret, and Jeff Clune. 2015. Neural Modularity Helps Organisms Evolve to Learn New Skills without Forgetting Old Skills. *PLoS Computational Biology* 11, 4 (04 2015), 1–24. https://doi.org/10.1371/journal.pcbi.1004128

[14] Josafath I. Espinosa-Ramos, Elisa Capecci, and Nikola Kasabov. 2019. A Computational Model of Neuroreceptor-Dependent Plasticity (NRDP) Based on Spiking Neural Networks. *IEEE Transactions on Cognitive and Developmental Systems* 11, 1 (3 2019), 63–72. https://doi.org/10.1109/TCDS.2017.2776863

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, Vol. 70. JMLR.org, 1126—-1135.

[16] W. Shane Grant, James Tanner, and Laurent Itti. 2017. Biologically Plausible Learning in Neural Networks with Modulatory Feedback. *Neural Networks* (2017). https://doi.org/10.1016/j.neunet.2017.01.007

[17] Jörg Hähner, Uwe Brinkschulte, Paul Lukowicz, Sanaz Mostaghim, Bernhard Sick, and Sven Tomforde. 2015. Runtime Self-Integration as Key Challenge for Mastering Interwoven Systems. In *Proceedings of the 28th International Conference on Architecture of Computing Systems*. VDE, 1–8.

[18] Albert W Hamood and Eve Marder. 2014. Animal-to-Animal Variability in Neuromodulation and Circuit Function. In *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 79. Cold Spring Harbor Laboratory Press, 21–28. https://doi.org/10.1101/sqb.2014.79.024828

[19] Gábor Herczeg, Tamás J Urszán, Stephanie Orf, Gergely Nagy, Alexander Kotrschal, and Niclas Kolm. 2019. Brain Size Predicts Behavioural Plasticity in Guppies (*Poecilia reticulata*): An Experiment. *Journal of Evolutionary Biology* 32, 3 (2019), 218–226. https://doi.org/10.5061/dryad.fp11572

[20] Jing Huang, Xiaogang Ruan, Naigong Yu, Qingwu Fan, Jiaming Li, and Jianxian Cai. 2016. A Cognitive Model Based on Neuromodulated Plasticity. *Computational Intelligence and Neuroscience* (2016). https://doi.org/10.1155/2016/4296356

[21] Khurram Javed and Martha White. 2019. Meta-Learning Representations for Continual Learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 1820–1830.

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming Catastrophic Forgetting in Neural Metworks. *Proceedings of the National Academy of Sciences of the United States of America* 114, 13 (mar 2017), 3521–3526. https://doi.org/10.1073/pnas.1611835114 arXiv:1612.00796

[23] Petr E Komers. 1997. Behavioural Plasticity in Variable Environments. *Canadian Journal of Zoology* 75, 2 (1997), 161–169.

[24] Jeffrey L. Krichmar. 2008. The Neuromodulatory System: A Framework for Survival and Adaptive Behavior in a Challenging World. *Adaptive Behavior* 16, 6 (2008), 385–399. https://doi.org/10.1177/1059712308095775

[25] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous Control with Deep Reinforcement Learning. In *ICLR (Poster)*.

[26] Nicolas Y. Masse, Gregory D. Grant, and David J. Freedman. 2018. Alleviating Catastrophic Forgetting using Context-Dependent Gating and Synaptic Stabilization. *Proceedings of the National Academy of Sciences* 115, 44 (2018), E10467–E10475. https://doi.org/10.1073/pnas.1803839115

[27] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory* (1989). https://doi.org/10.1016/S0079-7421(08)60536-8

[28] Frederic Mery and James G. Burns. 2010. Behavioural Plasticity: An Interaction between Evolution and Experience. *Evolutionary Ecology* 24, 3 (2010), 571–583. https://doi.org/10.1007/s10682-009-9336-y

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Amir Beattie, Charles Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236 arXiv:1312.5602

[30] Krzysztof Mogielski and Tadeusz Płatkowski. 2009. A Mechanism of Dynamical Interactions for Two-Person Social Dilemmas. *Journal of Theoretical Biology* (2009). https://doi.org/10.1016/j.jtbi.2009.06.007

[31] Rui Filipe Oliveira. 2012. Social Plasticity in Fish: Integrating Mechanisms and Function. *Journal of Fish Biology* 81, 7 (2012), 2127–2150. https://doi.org/10.1111/j.1095-8649.2012.03477.x

[32] Ting Qian, T. Florian Jaeger, and Richard Aslin. 2012. Learning to Represent a Multi-Context Environment: More than Detecting Changes. *Frontiers in Psychology* 3 (2012), 228. https://doi.org/10.3389/fpsyg.2012.00228

[33] Edward Robinson, Timothy Ellis, and Alastair Channon. 2007. Neuroevolution of Agents Capable of Reactive and Deliberative Behaviours in Novel and Dynamic Environments. In *Advances in Artificial Life*. Springer, 1–10. https://doi.org/10.1007/978-3-540-74913-4_35

[34] Tasmin L Rymer, Neville Pillay, and Carsten Schradin. 2013. Extinction or Survival? Behavioral Flexibility in Response to Environmental Change in the African Striped Mouse Rhabdomys. *Sustainability* 5, 1 (2013), 163–186. https://doi.org/10.3390/su5010163

[35] Emilie C. Snell-Rood. 2013. An Overview of the Evolutionary Causes and Consequences of Behavioural Plasticity. *Animal Behaviour* (2013). https://doi.org/10.1016/j.anbehav.2012.12.031

[36] Andrea Soltoggio, John A. Bullinaria, Claudio Mattiussi, Peter Dürr, and Dario Floreano. 2008. 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.*

[37] Judy A. Stamps. 2016. Individual Differences in Behavioural Plasticities. *Biological Reviews* (2016). https://doi.org/10.1111/brv.12186

[38] Kenneth O. Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. 2019. Designing Neural Networks through Neuroevolution. *Nature Machine Intelligence* 1 (2019), 25–35. https://doi.org/10.1038/s42256-018-0006-z

[39] Kenneth O. Stanley and Risto Miikkulainen. 2002. Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation* 10, 2 (2002), 99–127. https://doi.org/10.1162/106365602320169811

[40] Gido M. van de Ven, Hava T. Siegelmann, and Andreas S. Tolias. 2020. Brain-Inspired Replay for Continual Learning with Artificial Neural Networks. *Nature Communications* 11, 1 (2020). https://doi.org/10.1038/s41467-020-17866-2

[41] Nicolas Vecoven, Damien Ernst, Antoine Wehenkel, and Guillaume Drion. 2020. Introducing Neuromodulation in Deep Neural Networks to Learn Adaptive Behaviours. *PLOS ONE* 15, 1 (01 2020), 1–13. https://doi.org/10.1371/journal.pone.0227922

[42] Roby Velez and Jeff Clune. 2017. Diffusion-Based Neuromodulation can Eliminate Catastrophic Forgetting in Simple Neural Networks. *PLoS ONE* (2017). https://doi.org/10.1371/journal.pone.0187736

[43] Mark Viney and Anaid Diaz. 2012. Phenotypic Plasticity in Nematodes. *Worm* 1, 2 (2012), 98–106. https://doi.org/10.4161/worm.21086

[44] Simon X. Yang and Max Meng. 2000. An Efficient Neural Network Approach to Dynamic Robot Motion Planning. *Neural Networks* 13, 2 (2000), 143–148. https://doi.org/10.1016/S0893-6080(99)00103-3

[45] Simon X. Yang and Max Meng. 2000. An Efficient Neural Network Method for Real-Time Motion Planning with Safety Consideration. *Robotics and Autonomous Systems* 32 (2000), 115–128. https://doi.org/10.1016/S0921-8890(99)00113-X

[46] Xin Yao. 1999. Evolving Artificial Neural Networks. *Proc. IEEE* 87, 9 (9 1999), 1423–1447. https://doi.org/10.1109/5.784219

[47] B. W. Yap and C. H. Sim. 2011. Comparisons of Various Types of Normality Tests. *Journal of Statistical Computation and Simulation* 81, 12 (2011), 2141–2155. https://doi.org/10.1080/00949655.2010.520163

[48] Jason Yoder and Larry Yaeger. 2014. Evaluating Topological Models of Neuromodulation in Polyworld. In *Proceedings of the 14th International Conference on the Synthesis and Simulation of Living Systems*. MIT Press, 916–923. https://doi.org/10.7551/978-0-262-32621-6-ch149