Pain and Self-Preservation in Autonomous Robots: from Neurobiological Models to Psychiatric Disease.

Luca Piccolo[1,2], Fabio Dalla Libera[1], Andrea Bonarini[2], Ben Seymour[3,4], Hiroshi Ishiguro[1,4]

***Abstract*— The use of biologically realistic (brain-like) control systems in autonomous robots offers two potential benefits. For neuroscience, it may provide important insights into normal and abnormal control and decision-making in the brain, by testing whether the computational learning and decision rules proposed on the basis of simple laboratory experiments lead to effective and coherent behaviour in complex environments. For robotics, it may offer new insights into control system designs, for example in the context of threat avoidance and self-preservation. In the brain, learning and decision-making for rewards and punishments (such as pain) are thought to involve integrated systems for innate (Pavlovian) responding, habit-based learning, and goal-directed learning, and these systems have been shown to be well-described by RL models. Here, we simulated this 3-system control hierarchy (in which the innate system is derived from an evolutionary learning model), and show that it reliably achieves successful performance in a dynamic predator-avoidance task. Furthermore, we show situations in which a 3-system architecture provides clear advantages over single or dual system architectures. Finally, we show that simulating a computational model of obsessive compulsive disorder, an example of a disease thought to involve a specific deficit in the integration of habit-based and goal-directed systems, can reproduce the results of human clinical experiments. The results illustrate how robotics can provide a valuable platform to test the validity and utility of computational models of human behaviour, in both health and disease. They also illustrate how bio-inspired control systems might usefully inform self-preservative behaviour in autonomous robots, both in normal and malfunctioning situations.**

I. INTRODUCTION

Progress in the design of bio-inspired control systems for autonomous robots have illustrated two important differences when compared to real biological agents. First, the behavior of animals is strongly governed by their sense of self-preservation, since first and foremost they act to defend themselves against dangerous and mortal threats before engaging in otherwise routine reward harvesting [1]. Second, animals have an internal hierarchy of control systems, from highly automatic responding to complex deliberative reasoning, which has evolved to balance the complexity of real-world environments with computational cost [2]. The prospect of embracing these characteristics offers the opportunity both to improve control systems for autonomous robots, and enhance their observable biological realism.

Reciprocally, insights from control theory and robotics have inspired an understanding of the nature of action learning and control in the brain. In particular, Reinforcement Learning (RL) has proved a valuable framework for understanding learning from experience in animal and human experiments, including both state-dependent (Pavlovian) and action-dependent (instrumental) learning [3]. However, these models usually consider highly simplified decision-making problems, and it remains unclear whether those that have emerged for different levels of control can be integrated together such that they are capable of coherent and effective control on the sorts of test-beds used in robotics (although there are existing computational accounts of integration of model-based and model free systems [4]–[6] and innate-model-free [7], [8], there are few accounts of 3-system integration). As well as being a key test of their validity, this is important because recent theories of human psychiatric disease have proposed quite specific deficits in their computational architecture, especially in models of punishment learning [9], [10].

With these dual perspectives in mind, we aimed to test neurobiological models of learning and action control within a simulated robotics framework, and explore their application to computational models of aberrant control. Specifically, we set out three core aims:

- To consolidate computational neurobiological models of a 3-system hierarchy of control incorporating innate (Pavlovian) responding, habitual actions (model-free control), and goal-directed (model-based control) actions from a RL perspective, with particular emphasis on punishment (escape and avoidance) learning.
- To test whether implementation of this integrated model could yield effective avoidance behaviour in a dynamic predator task, and whether the incorporation of 3 systems harbours clear advantages to single-system controllers in a number of specific situations.
- To explore whether disease specific hypotheses can be modeled within this framework, taking the example of obsessive compulsive disorder (OCD) as a deficit in the transition of control between habit and goal-directed controllers [11].

[1]LP, FDL and HI are at the Department of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531luca.piccolo2@outlook.com, fabio.dl@irl.sys.es.osaka-u.ac.jp, ishiguro@sys.es.osaka-u.ac.jp

[2], LP and AB are at the AI and Robotics Lab, Department of Electronics, Information, and Bioengineering, Politecnico di Milano andrea.bonarini@polimi.it

[3]BS is at the Computational and Biological Learning lab, Department of Engineering, Trumpington Street, Cambridge University, CB2 1PZ, UK. bjs49@cam.ac.uk

[4]BS and HI are at the Center for Information and Neural Networks, NICT, 1-4 Yamadaoka, Suita City, Osaka, 565-0871. BS and HI contributed equally to this work.

BS is funded by the Wellcome Trust (UK), the National Institute for Information and Neural Networks (Japan), and Arthritis Research UK. HI is funded by Grant-in-Aid for Scientific Research (S) No. A252200040.

Achieving these aims would illustrate both the utility of a robotics framework in understanding human decision-making in health and disease, and may enhance the realism of bio-inspired robots, especially in the domain of defensive and self-preservative behaviour.

II. An Integrated Computational Model of Human Control Systems

The notion that animal and human action control is supported by more than one system has a long history in psychology [12]. The conventional dual system ('automatic' and 'deliberative') account has given way to an recognition that there are 3 core systems for action control: innate/Pavlovian responding, in which state-dependent values yield hard-wired responses such as approach and withdrawal; habit-based responding in which action values are learned through a simple process of reinforcement or suppression; and goal-directed responding involving potentially complex internal representations of state and action space, incorporating transition probabilities and the identity of reward and punishment outcomes [13]. Recent human behavioural and neuroimaging studies have provided critical insight into the underlying computations involved in these systems. Based on original work on dopamine neuron responses in monkeys, there is good evidence that Pavlovian learning for both rewards [14], [15] and punishments [16] is well characterized by RL (temporal difference) models. In the brain, this involves mutually inhibitory prediction error responses for reward and punishment converging in the ventral striatum [17]. Simple (presumptively habit-based) action learning has also now been well studied, with good evidence of RL-like prediction errors in dorsal regions of striatum [18], [19]. In the case of punishment (pain), positively-valenced avoidance prediction errors are also seen in dorsal striatum [20], although negatively-valenced prediction errors have not been consistently identified [21], raising a debate about whether avoidance is achieved primarily by treating punishments as negative rewards with action being governed by a reward system incentivized by safety-states [22]. This structure is based on classical two-factor theories of avoidance learning, which posit that the inhibitory *state* of safety (i.e. predicting the absence of punishment) motivates instrumental *action* - an architecture that has close parallels with actor-critic models of learning [23].

RL models of goal-directed behavior have been based on experimental paradigms with either complex state-action transition structures or abstract rules that govern outcome probabilities [24]. But at least for rewards, there is evidence of error-based brain responses sensitive to these structural internal models that cannot be accounted for by simple reinforcement [25]. Again, these responses appear to converge on the dorsal striatum, although possibly in more medial regions than habit-based prediction errors [26]. Algorithmically, there is in principle considerable scope for a diversity of representations, ranging from simple representations of state-transitions [27], [28], spatial maps [29] to potentially complex decision-trees that can support dynamic planning.

With several controllers capable of directing behavior, more recent attention has been paid to exactly how the brain decides how to choose between or integrate the decisions of each system. In the case or Pavlovian responses, it may be that they directly compete with or bias instrumental actions, effectively providing a rapid, innate impulse to do something, which may be especially important in the case of threats [7], [30]. With regards to interactions between habit-based and goal-directed controllers, some more sophisticated theories propose some sort of uncertainty-based competition between the outputs of each [4], [5]. Note, we refer to the innate system as 'Pavlovian' since Pavlovian learning involves performing innately-specified behavioural programs, although innate responses are emitted in the absence of Pavlovian learning, when directly presented with an aversive or appetitive (unconditioned) stimulus. Also, although Pavlovian learning can be model-based [31], [32], we model it here as a model-free process.

A summary of the architecture of a 3 system model for integrated reward and punishment behavior is illustrated in Figure 1. The model provides an architecture for control based on the key results from human experiments. For simplicity, it omits many additional details of control, including forms of Pavlovian-instrumental interaction (conditioned-suppression, Pavlovian-instrumental transfer), risk behaviour (risk and ambiguity aversion), and different types of conditioned response (preparatory and consummatory).

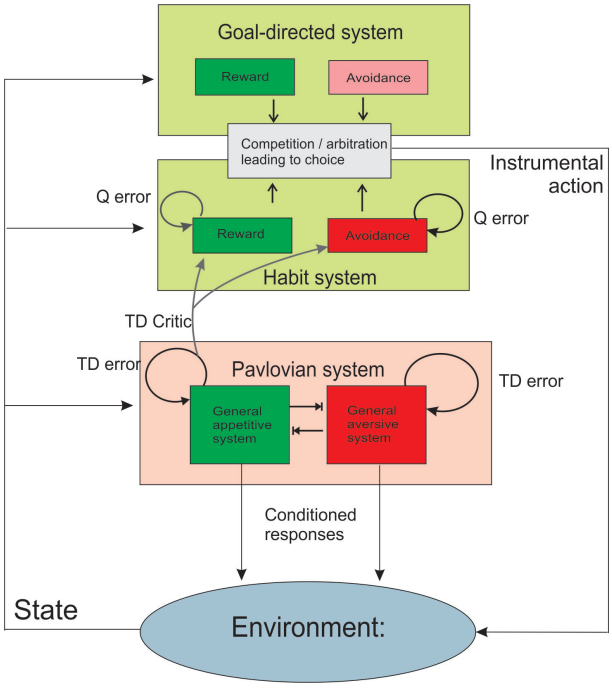

Fig. 1. **Schematic diagram of integrated control systems for human reward and punishment learning** An agent-environment control diagram showing 3 levels of control. Pavlovian learning reflects state-based learning, but emits conditioned responses that may or may not influence the environment. The habit-based action system achieve avoidance using a reward-based critic that derives from inhibition between reward and punishment systems. Habit and goal-directed action systems compete for control following some sort of arbitration interaction

III. Computational model

Following the neurobiological data, we set out to model behaviour as determined by the joint control of three components: innate (which we hereafter refer to as Pavlovian) responses, model-free (habit-based) control and model-based (goal-directed) control, to study their effect on the overall system behavior. We simulated a prey in a predator-prey setting in a virtual world, and observed how different combination of modules affect the capability of the prey in escaping. Section III-A provides more details on the simulated virtual world. Successively, the Pavlovian, model-free and model based subsystem are described in Sections III-B, III-C and III-D, respectively. Finally, Section III-E describes how the modules are integrated to actually control the prey agent.

A. Simulated environment

In our experiments a simple environment, where the effect of the three subsystems can be clearly observed, was preferred over more realistic predator-prey simulation. Specifically, a time discrete simulation in a discretized space was employed. The simulated world, shown in Fig. 2, comprised obstacles and free areas in which the simulated predator and prey agents could move.

The predator behavior was fixed (i.e. we assume the prey is the only agent for which learning occurs), to remove possible confounding effects caused by their co-evolution. The predator always proceeded toward the prey in an (approximately) straight line. The prey could move two times for each movement of the predator, so that escaping was possible. Simulation was run over a set of episodes. At the beginning of each episode, the agents were placed at a random (free) position satisfying two conditions: the distance between the two agents was within a predefined range (half of the map side length) and the predator and the prey were in sight (i.e. no obstacle blocks were placed between them). Every time the distance between the two agents exceeded a certain threshold (halt of the map side length), or when the prey hid behind an obstacle, a new episode was started.

The fear/pain perceived by the prey was assumed to be a function of the distance to the predator, with "stronger" fear was associated with shorter distances. Specifically, denoting by d the prey-predator distance and by l the map side length, the reward r was set as -5 for $d < 0.1l$, -4 for $0.1l \leq d < 0.25l$, -3 for $0.25l \leq d < 0.3l$, -2 for $0.3l \leq d < 0.4l$, -1 for $0.4l \leq d < 0.5l$ and 0 otherwise.

B. Pavlovian responses

Pavlovian responses typically reflect simple approach/withdrawal responses, and were simulated using a feed forward network with a single hidden layer with sigmoidal activation function. The input of the network was a set of categorical variables describing the surrounding of the prey. The output was the direction of movement.

In order to provide a sufficiently realistic simulation, the hexagonal tessellation used for describing the environment and the agents position is relatively fine. However, to produce meaningful behavior, the agents must be aware of a sufficiently wide surrounding area. For this reason, a coarser subdivision of the world was introduced.

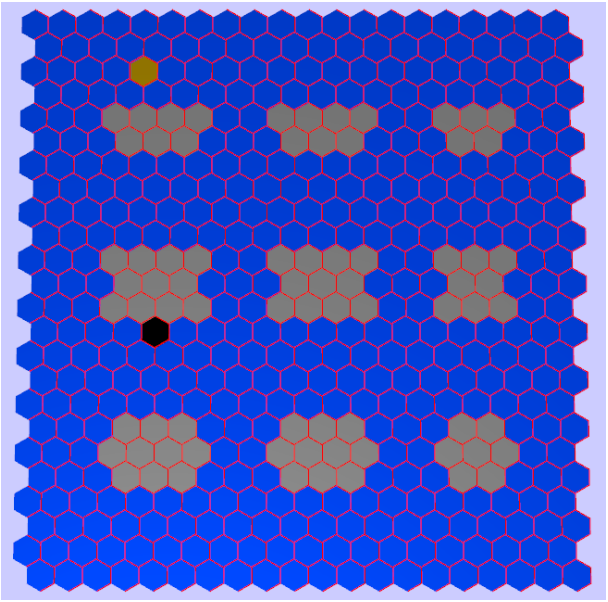

Fig. 2. **Map used for the experiments**. The simulated environment consisted of an approximately square area tessellated using regular hexagons (this kind of tessellation presents the important property of assuring the same Euclidean distance between the centers of all the tiles that have a common vertex). In this figure, color indicates the content of the tile: blue for free areas, gray for obstacles, yellow for the pray and black for the predator.

Specifically, the whole space was divided into 49 areas, where each area $A_{i,j}$, $1 \leq i,j \leq 7$, consisted of all the hexagonal tiles for which the center c_x, c_y has coordinates satisfying $l\frac{i-1}{7} \leq c_x \leq l\frac{i}{7}$ and $l\frac{j-1}{7} \leq c_y \leq l\frac{j}{7}$. Each area was described by a categorical variable taking one of the following values: *obstacle*, *predator* or *empty*. An area is categorized as *obstacle* when at least 20% of its tiles consist of an obstacle, as *predator* if the predator's tile is contained in the area, and *empty* otherwise. The categories of the area containing the prey and of the 8 neighboring areas (in other terms, the categories of the prey's Moore neighborhood), represented with a 1-of-c coding, constituted the neural network's input. The network's output consisted of 9 neurons, one for each of these surrounding areas. The center of the area corresponding to the most strongly activated neuron was identified, and the movement in the discrete world that best approximates a movement toward such center was used as the Pavlovian response.

Pavlovian responses reflect an inherited set of primitive actions that are learned through evolution. To capture this, we set the weights of the network by neuroevolution. The genetic algorithm was set as follows:

- population size of 500 individuals evolved for 15000 generations,
- value encoding, initial genome drawn from a uniform distribution in the range $[-0.1, 0.1]$,
- single point crossover with probability 0.9,
- floating point additive mutation, probability of mutation for each gene 0.1, mutation increment sampled from a

uniform distribution in the range $[-0.2, 0.2]$,
- roulette wheel selection, with elitism (4 individuals).

Care was placed in the gene's ordering. Specifically in the array we first placed all the weights of the arcs going into the first hidden neuron, then all the weights of the arcs going out from that same neuron, then the arcs going in the second hidden neuron, etc. In this way, the crossover operator is able to maintain part of the "features" (nonlinear combinations of the inputs) and their effect on the output probabilities computed by an individual. The same property does not hold if the weights are simply ordered by layer.

The evaluation function of genetic algorithm's chromosome consisted of the average reward[1] obtained by a prey controlled by the corresponding neural network over a set of simulations. Specifically, each neural network was tested over 28 episodes, each comprising 80 time steps. The initial locations of the agents for each of the 28 episodes was decided beforehand and kept constant over the evaluations to decrease the variance in the agent's evaluations. We experimentally confirmed that no overfitting emerged from this choice.

As a final note, it is worth noting that additional layers did not show performance improvements, while making the mapping fully linear (i.e. removing the hidden layer) was found to be strongly detrimental.

C. Model-free subsystem

Model-free decision was modeled using an actor-critic model. Using the notation of [33]:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{1}$$

$$V(s_t) \leftarrow V(s_t) + \alpha * \delta_t \tag{2}$$

$$\pi(s_t, a_t) \leftarrow \pi(s_t, a_t) + \beta * \delta_t \tag{3}$$

where π are the modifiable policy parameters of the actor, $\alpha = 0.02$, $\beta = 0.02$, and $\gamma = 0.99$.

The state comprised three components:

1) The area (defined as explained in Section III-B) was encoded as a single number. Using a single hex tile would have lead to an excessively big state space, and thus areas were used.
2) The direction from which the predator was approaching, expressed as the angle of the segment linking the two agents, discretized over 8 values.
3) The distance between the predator and the prey, discretized over three values: below 1/10 of the map size, in the range 1/10 to 1/4 of the map size, and over 1/4 of the map size.

As done for the Pavlovian system, nine possible actions were defined. These corresponded to the movement toward the center of one of the areas in the Moore neighbourhood of the current prey area. The actor employed an ϵ-greedy greedy policy: with probability ϵ ($\epsilon = 0.1$ in our implementation) the action is random, and with probability $1 - \epsilon$ it is $\arg\max_a p(s_t, a)$.

[1]Here and in the following we use the term reward, but it should be noted that rewards are always negative or null, as they express the fear/pain level perceived.

D. Model-based subsystem

The model-based subsystem evaluates the goodness of each action a_t by performing a series of K Monte Carlo simulations starting at the current state s_t and performing a_t as the first action. The action that, on average, yields the highest reward is then chosen.

To provide more detail, states and actions were the same as utilised for the model-free subsystem described in Section III-C. The models for the state-to-state transition probability and the rewards were obtained by the agent through online experience. For each start-state action pair, the last L (in our implementation 4) states reached and rewards received are stored. These L experiences are taken as an approximation of the expected immediate reward and expected transition probability. Keeping a history of events of limited size, besides providing a constant upper limit on the computational resources, assures the model is able to quickly reflect changes in the environment.

The policy used during the Monte Carlo simulation was a simple ϵ-greedy one. With probability ϵ_M the action is random, and with probability $1 - \epsilon_M$ it chooses the action whose expected immediate reward (the average of the history for the current state and that action) is the highest. Note that the reward for state-action pairs never explored is assumed to be 0, i.e. the highest possible value, since rewards are only negative (see Section III-A). This is used to favour exploration in the early stages of learning.

The rewards obtained during a fixed number of steps $0 \leq i < M$ using this greedy policy, discounted by a factor γ_M^i, are summed to obtain the evaluation for action a for a single Monte Carlo run. As previously mentioned, the average over K Monte Carlo simulation is taken as an indication of the value of action a. The action leading to the highest value is chosen with probability $1 - \epsilon_G$, while a random action is output with probability ϵ_G. In our implementation $M = 5$, $K = 10$, $\epsilon_M = \epsilon_G = 0.1$, $\gamma_M = 0.99$.

We conclude with a remark on the choice of this kind of simple implementation for the model-based subsystem. One of the main requirements for this module is that it should be very reactive in the face of change with respect to the model-free module. One possibility could be to use $Q(\lambda)$ (see [34]), but such an approach would not be biologically faithful because the computational load would be distributed among all time steps. A similar reasoning led to the exclusion of Dyna-PI and Dyna-Q (see [35] and [36], respectively). A good, biologically plausible candidate would be the algorithm presented in [5], but without modifications with the state and action spaces that we are currently using the learning times are prohibitive, especially when compared to the model-free module.

E. Arbitration

The idea behind the integration of the modules is that for strong and sudden fear/pain the module determining

the behavior should be the simplest and fastest, that is the Pavlovian subsystem. When the environment is easily predictable, in other terms there is no "surprise", the habit, i.e. the model-free system should become a reliable provider of control. Conversely, when the environment yields frequent unexpected changes, more substantive goal-based reasoning is required and the model-based subsystem should be prioritized.

The value of δ can be taken as an indication of the "surprise", while the reward r itself constitutes a measure of the fear in our implementation. In practice, the value of δ has a dynamic that is too fast for arbitration, so a low pass filtering is opportune. Furthermore, as it may be desirable to model independently positive surprises (in our case the absence of expected pain) and negative surprises (unexpected pain), we define the two following quantities:

$$\begin{cases} \bar{\delta}_{pos} \leftarrow (1-\rho) * \bar{\delta}_{pos} + \rho * \max(\delta, 0) \\ \bar{\delta}_{neg} \leftarrow (1-\rho) * \bar{\delta}_{neg} + \rho * \min(\delta, 0) \end{cases} \tag{4}$$

correspondingly to positive and negative surprises, respectively ($\rho = 0.1$ in our experiments).

With the above definitions, the arbitration becomes:

$$\begin{array}{l} \textbf{if } (\bar{\delta}_{neg} < \kappa_{pavl\delta neg} \textbf{ AND } last\_reward < \kappa_{pavlr}) \\ \quad use\ Pavlovian \\ \textbf{else if } (\kappa_{modbneg} < \bar{\delta}_{neg} \textbf{ OR } \bar{\delta}_{pos} > \kappa_{modbpos}) \\ \quad use\ model\_based \\ \textbf{else} \\ \quad use\ model\_free \end{array} \tag{5}$$

where $\kappa_{pavl\delta}$, κ_{pavlr}, $\kappa_{modbneg}$ and $\kappa_{modbpos}$ are opportune constants, with $\kappa_{pavl\delta neg} < \kappa_{modbneg}$ (in our implementation $\kappa_{pavl\delta} = -7$, $\kappa_{pavlr} = -3$, $\kappa_{modbneg} = -3$ and $\kappa_{modbpos} = 1$).

Other models for the integration of model-based and model-free subsystems, based on the relative uncertainty of each subsystem, were recently proposed [5], [27]. In [27] the uncertainty, defined as SPE, is computed, however the actual arbitration between the modules is a sole function of the elapsed time (an exponential decay), and thus not suited (neither biologically plausible) for our setup. Conversely, the approach presented in [5] could be extended to fit our setup, and comparison of our simple model with an integration derived from Lee et al.'s work is an important topic for future work. A fundamental difference between the arbitration module described in this Section and the ones based on relative uncertainty is the stage at which arbitration can occur. In our case, arbitration can occur either before or after the computation of the optimal actions by each module. When arbitration is done beforehand, unselected modules can skip their computation. In case an action and its relative uncertainty is needed, however, the modules considered for integration need to compute their solutions, with relative uncertainty, and only afterwards the arbitration can choose which module's action to execute.

IV. Experiments

The first goal of the experimental simulation was to confirm that each of the three modules, when used alone, is able to produce useful behaviors. The second goal was to explore whether they could be integrated, and to observe the influence that each component brought to the overall system. Finally, as an example of the capability of the model in simulating different emergent behaviors by simple alteration of the parameter values, we aimed to show how an experimental OCD-like behavior could be reproduced.

Figure 3 shows that learning converges for all three modules, when applied alone. It is important to note that the neuroevolutionary (i.e. the genetic algorithm) used for the Pavlovian module was carried out only once, before the integration of the modules, and then the weights are fixed. This corresponds to the evolution among generations of individuals. Once an individual is born, its hard-coded responses are immutable. On the other hand, model-free and model-based subsystems are assumed to model the learning that occurs in the single individual. It could be argued that in real animals evolution of the Pavlovian response can continue after model-free and model-based reactions appears, but for simplicity we opted for a fixed set of responses.

We also note that the model-based module is designed to be very reactive (only the last 4 state-action-reward sets are stored) and its implementation is stochastic (10 Monte Carlo simulations), thus it has much more variability in the chosen action (and consequently, on the rewards achieved) compared to the other modules.

Figure 4 provides a comparison on the model-free module acting alone and when it acts together with the Pavlovian responses. It can be seen that even if in the long run the Pavlovian subsystem slightly decreases the performance of the model-free module acting alone, in the initial phases of the learning adding Pavlovian responses is advantageous.

A similar effect can be observed for the combination of model-free and model-based learning shown in Fig. 5. Again, the model-based module gives an initial advantage, at the expense of a later decrease of the overall performance.

Finally, Fig. 6 shows the effect of adding the model-based system (which one assumes requires some higher-order cognition) to the combination of Pavlovian and model-free subsystem (which is likely to be present even in primitive animals). The addition of the model-based subsystem provides a cost over long term performance, but with an advantage in early learning (difficult to see in the figure because of the steepness of the curve). This is more clear in Table I, which reports the performance increase for the first e episodes, for various values of e, obtained by adding the goal-based subsystem.

Figure 7 shows how frequently each module is activated as learning goes on. It clearly illustrates how the model-based and Pavlovian responses influence the initial stages of learning, giving more and more control to the model-free system over time.

In the last experiment, we aimed to simulate a characteristic experimental (pathological) behaviour observed as a hallmark of Obsessive-Compulsive Disorder (OCD). According to a novel behavioural neuroscientific theory [11], OCD is proposed to emerge as an enhanced reliance on model-

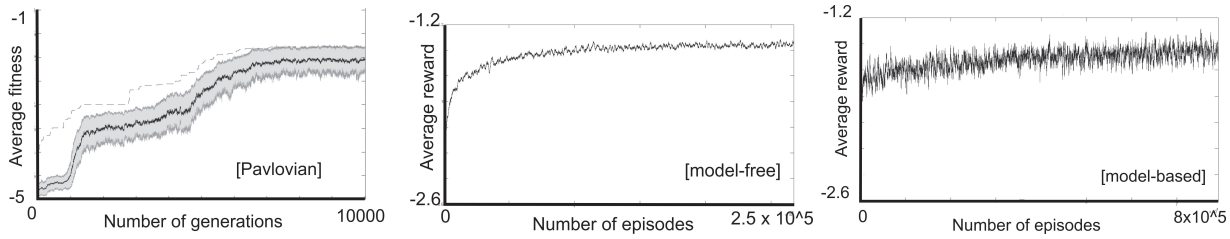

Fig. 3. **Learning convergence of the single modules**. The left panel shows the convergence of the neuroevolution for the Pavlovian (innate) module. The convergence of the model-free module is shown in the central panel. The convergence of the model-based module in reported in the right panel. Note that here and in the following figures a moving average filter of width 1000 is applied on the curves to make them more legible.

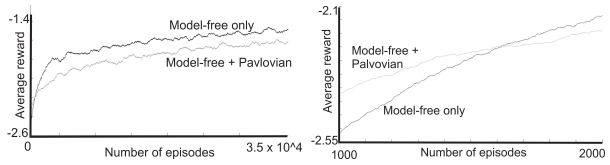

Fig. 4. **Combination of Pavlovian responses and model-free subsystem**. The left panel shows a comparison between the model-free module acting alone (darker curve) and when acting together with the Pavlovian subsystem (lighter curve). The right panel shows with higher detail the initial part of the curves, highlighting the advantage brought by the Pavlovian reactions.

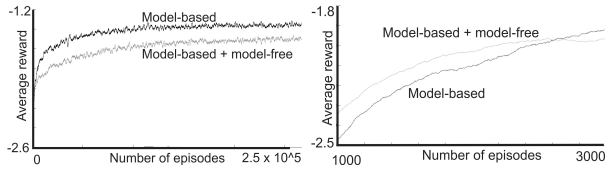

Fig. 5. **Combination of the model-free and model-based subsystem**. The left panel shows a comparison between the model-free module acting alone (darker curve) and when acting together with the model-based subsystem (lighter curve). The right panel shows with higher detail the initial part of the curves, highlighting the advantage brought by the model-based subsystem.

free habits during avoidance. In a recent experiment, OCD patients were trained to perform an action to avoid a punishment, they observed a relative failure of subjects to shift from model-free to model-based control when punishments were devalued (i.e. when an outcome that was previously punishing was no longer so, because of an experimental rule informed to the patients) action [37]. This fit with the hypothesis that this derives from a reduced transition from model-free control to model-based control i.e. excessive avoidance 'habits'.

e	Average reward increase
100	+0.1375
200	+0.1723
300	+0.1891
500	+0.1772
750	+0.1565
1000	+0.1283

TABLE I

ADVANTAGE OF ADDING THE GOAL-BASED SUBSYSTEM TO THE OTHER TWO MODULES, OBTAINED FOR THE FIRST e EPISODES, AVERAGED OVER 10 INDEPENDENT RUNS.

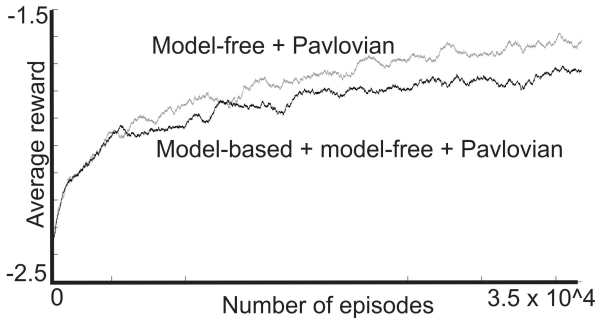

Fig. 6. **Combination of the three subsystems**. Comparison between the combination of model-free and Pavlovian subsystem (lighter curve) with respect to the three modules acting together (darker curve). Over a long horizon, it can be seen that the combination of 3 systems is disadvantageous, but over the first few hundred trials, there is a clear advantage, as outlined in Table 1.

We aimed to mimic this result using our model. First, the prey underwent the usual learning in presence of the predator causing pain/fear. When learning converged, the sign of the rewards was reversed, such that the predator delivered a positive reward. In this case, thus, staying still (or even better proceeding toward the predator) clearly becomes the best policy.

We tested the behavior of two versions of the prey: a "healthy individual" and an "OCD patient" version. In order to simulate the impairment in the activation of model-based reasoning, the value of $\kappa_{modbpos}$ was raised for the "OCD patient" prey, and left unvaried for the "healthy individual".

The behavior of the two preys is compared in Fig. 8. The vertical axis reports the time required for a prey to have a distance for the predator exceeding 1/4 of the map size, or hiding behind an obstacle. Clearly, this time should be the shortest possible during normal training, and the longest possible (theoretically, infinite) during the phase with reverse rewards. We found that the "healthy individual" was much quicker at increasing this time after the rewards inversion, exactly as observed in real experimental patients.

V. DISCUSSION

The experiment illustrates that a neurobiologically based 3-system control model of learning and decision-making can support robust control in simulated autonomous agents in complex environments. We showed that the integration of different controllers does not disrupt the convergence

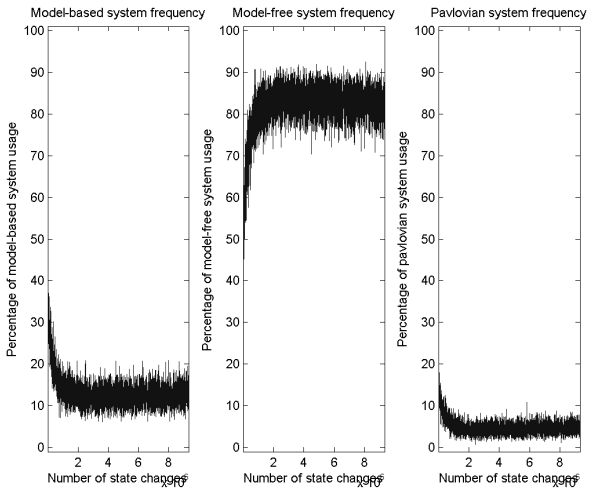

Fig. 7. **Frequency of activation of the three modules.**. From the left to the right we have the percentages of activation during a sample run of the three modules model-based, model-free and Pavlovian. The percentages were obtained passing through a moving average filter of size 5000 the three vectors containing the integer 1 when each module was activated and 0 when it was not, then multiplying the results for 100.

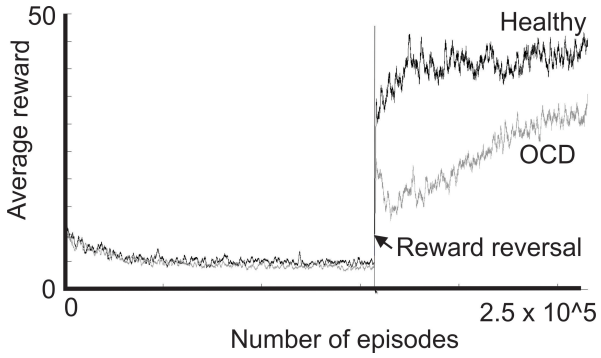

Fig. 8. **Comparison between healthy subject and OCD patient**. The time taken by the preys to escape (i.e. having a distance for the predator exceeding 1/4 of the map size, or hiding behind an obstacle) for an "OCD patient" (lighter curve) and an "healthy individual"(darker curve).

of either of them, and that each system contributes an important advantage to control. That is, Pavlovian responses and model-based reasoning provide an initial advantage in new environments, at the expense of a lower maximum performance when learning converges. We also showed that this framework can be used to test computational theories of psychiatric disease, taking the specific case of OCD as an example.

The results illustrates how an integrated system prioritizes short-term safety over long term performance, which is likely to be especially important in uncertain, dynamic and dangerous environments. This illustrates a novel strategy for 'Safe RL' [38]. Note that the model-based system and the innate/Pavlovian system acheive this in different ways: the model-based system by rapid, computationally expensive new learning of contingencies in the face of uncertainty, and the innate/Pavlovian system by importing the wisdom of evolutionary knowledge as a sort of 'prior' on the action value space.

From the perspective of neuroscience, these findings are useful because existing computational theories have been developed and tested on highly simplified tasks, such as one-step or two-step bandit tasks with a small set of available actions. But scaling these models up and allowing their integration and interaction is not trivial, and their capacity to support robust behaviour in complex, dynamic environments has not previously been tested. It therefore provides an important demonstration of their validity.

Inevitably with a complex, multi-controller model of control, the parameter space is large, and there are several assumptions and approximations required to take individual components of neurobiological models into an integrated systems-level model. It is beyond the scope of this paper to exhaustively consider how model performance depends on parameter values, but using computationally and biologically reasonable parameter values shown to lead to clear results. Another area of uncertainty in the literature to date is on the integration and arbitration of controllers [4], [5]. But despite this complexity and uncertainty, the robustness of the architecture to parameter changes in our experiments provides support to the the validity of the human/animal multi-system architecture. Indeed, parameter tuning may allow significant further optimisation, for instance in regards to the long-term reduction in performance that results from having Pavlovian and model-based controllers in stable, largely stationary environments (to whatever extent this is or is not a realistic circumstance).

Of particular interest is the applicability of the modelling approach presented to understanding disorders in neuroscience. There is now a growing argument that several psychiatric disorders can be understood in terms the dysfunction of specific computational mechanisms - a field called computational psychiatry [9], [39]. In the context of aversive behaviours, this includes disorders such as OCD, phobias, post-traumatic stress disorder, anxiety disorder, depression, and chronic pain. However, despite plausible computational hypotheses for each, there is still a large step to be traversed between a simple model and a complex behavioural phenotype. In our study of OCD here, although we have not modelled a full phenotype of compulsive behaviours clinically observed, we at least illustrate the capacity of our model to accommodate what is thought to be one of the most characteristic experimental findings observed - the over-reliance on avoidance habits. This demonstration is intended to illustrate the principle of a constructivist approach to disease modelling, as opposed to provide anything like a comprehensive account of OCD. As such, we hope that this framework could be used to characterise a fuller set of experimental behaviours and naturalistic 'symptoms' in OCD and other disorders.

The model we present also holds useful insights into control systems for robots, since we show that multi-component control systems can convey clear advantages in certain situations, especially early learning. In so doing, they may also help achieve a separate goal in robotics

- to enhance life-likeness of robots. In particular, it is likely that life-likeness might be especially enhanced not so much by extraordinary computational capabilities of robots, but by their assimilation of human imperfections, such as proneness for errors, behavioural traits such as impulsivity and compulsivity, and susceptibility to psychiatric disease. This latter point also illustrates the capacity for autonomous robots to develop their own 'psychiatric' malfunction when adopting bio-inspired architectures.

REFERENCES

[1] K. Doya and E. Uchibe, "The cyber rodent project: Exploration of adaptive mechanisms for self-preservation and self-reproduction," *Adaptive Behavior*, vol. 13, no. 2, pp. 149–160, 2005.

[2] A. R. Otto, S. J. Gershman, A. B. Markman, and N. D. Daw, "The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive," *Psychological science*, vol. 24, no. 5, pp. 751–761, 2013.

[3] P. Dayan and N. D. Daw, "Decision theory, reinforcement learning, and the brain," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 8, no. 4, pp. 429–453, 2008.

[4] N. D. Daw, Y. Niv, and P. Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nature neuroscience*, vol. 8, no. 12, pp. 1704–1711, 2005.

[5] S. W. Lee, S. Shimojo, and J. P. ODoherty, "Neural computations underlying arbitration between model-based and model-free learning," *Neuron*, vol. 81, no. 3, pp. 687–699, 2014.

[6] M. Khamassi and M. D. Humphries, "Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies," *Frontiers in behavioral neuroscience*, vol. 6, 2012.

[7] P. Dayan, Y. Niv, B. Seymour, and N. D. Daw, "The misbehavior of value and the discipline of the will," *Neural networks*, vol. 19, no. 8, pp. 1153–1160, 2006.

[8] P. Dayan and K. C. Berridge, "Model-based and model-free pavlovian reward learning: revaluation, revision and revelation," *Cognitive, affective & behavioral neuroscience*, vol. 14, no. 2, p. 473, 2014.

[9] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, "Computational psychiatry," *Trends in cognitive sciences*, vol. 16, no. 1, pp. 72–80, 2012.

[10] Q. J. Huys, T. V. Maia, and M. J. Frank, "Computational psychiatry as a bridge from neuroscience to clinical applications," *Nature neuroscience*, vol. 19, no. 3, pp. 404–413, 2016.

[11] C. M. Gillan, M. Papmeyer, S. Morein-Zamir, B. J. Sahakian, N. A. Fineberg, T. W. Robbins, and S. de Wit, "Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder," *American Journal of Psychiatry*, vol. 168, no. 7, pp. 718–726, 2011.

[12] J. S. B. Evans, "In two minds: dual-process accounts of reasoning," *Trends in cognitive sciences*, vol. 7, no. 10, pp. 454–459, 2003.

[13] A. Dickinson and B. Balleine, "Motivational control of goal-directed action," *Animal Learning & Behavior*, vol. 22, no. 1, pp. 1–18, 1994.

[14] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[15] J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan, "Temporal difference models and reward-related learning in the human brain," *Neuron*, vol. 38, no. 2, pp. 329–337, 2003.

[16] B. Seymour, J. P. O'doherty, P. Dayan, M. Koltzenburg, A. K. Jones, R. J. Dolan, K. J. Friston, and R. S. Frackowiak, "Temporal difference models describe higher-order learning in humans," *Nature*, vol. 429, no. 6992, pp. 664–667, 2004.

[17] B. Seymour, J. P. O'doherty, M. Koltzenburg, K. Wiech, R. Frackowiak, K. Friston, and R. Dolan, "Opponent appetitive-aversive neural processes underlie predictive learning of pain relief," *Nature neuroscience*, vol. 8, no. 9, pp. 1234–1240, 2005.

[18] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, "Representation of action-specific reward values in the striatum," *Science*, vol. 310, no. 5752, pp. 1337–1340, 2005.

[19] J. O'doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan, "Dissociable roles of ventral and dorsal striatum in instrumental conditioning," *science*, vol. 304, no. 5669, pp. 452–454, 2004.

[20] B. Seymour, N. D. Daw, J. P. Roiser, P. Dayan, and R. Dolan, "Serotonin selectively modulates reward value in human decision-making," *Journal of Neuroscience*, vol. 32, no. 17, pp. 5833–5842, 2012.

[21] E. Eldar, T. U. Hauser, P. Dayan, and R. J. Dolan, "Striatal structure and function predict individual biases in learning to avoid pain," *Proceedings of the National Academy of Sciences*, vol. 113, no. 17, pp. 4812–4817, 2016.

[22] J. A. Dinsmoor, "Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing," *Journal of the experimental analysis of behavior*, vol. 75, no. 3, pp. 311–333, 2001.

[23] T. V. Maia, "Two-factor theory, the actor-critic model, and conditioned avoidance," *Learning & behavior*, vol. 38, no. 1, pp. 50–67, 2010.

[24] A. N. Hampton, P. Bossaerts, and J. P. Odoherty, "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans," *Journal of Neuroscience*, vol. 26, no. 32, pp. 8360–8367, 2006.

[25] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan, "Model-based influences on humans' choices and striatal prediction errors," *Neuron*, vol. 69, no. 6, pp. 1204–1215, 2011.

[26] L. A. Bradfield and B. W. Balleine, "Thalamic control of dorsomedial striatum regulates internal state to guide goal-directed action selection," *Journal of Neuroscience*, pp. 3860–16, 2017.

[27] J. Gläscher, N. Daw, P. Dayan, and J. P. O'Doherty, "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning," *Neuron*, vol. 66, no. 4, pp. 585–595, 2010.

[28] E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, and N. D. Daw, "Predictive representations can link model-based reinforcement learning to model-free mechanisms," *bioRxiv*, p. 083857, 2016.

[29] D. A. Simon and N. D. Daw, "Neural correlates of forward planning in a spatial decision task in humans," *Journal of Neuroscience*, vol. 31, no. 14, pp. 5526–5539, 2011.

[30] P. Dayan and B. Seymour, "Values and actions in aversion," *Neuroeconomics: Decision making and the brain*, pp. 175–191, 2008.

[31] C. Prévost, D. McNamee, R. K. Jessup, P. Bossaerts, and J. P. O'Doherty, "Evidence for model-based computations in the human amygdala during pavlovian conditioning," *PLoS computational biology*, vol. 9, no. 2, p. e1002918, 2013.

[32] A. Jauffret, N. Cuperlier, P. Tarroux, and P. Gaussier, "From self-assessment to frustration, a small step toward autonomy in robotic navigation," *Frontiers in neurorobotics*, vol. 7, 2013.

[33] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[34] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, 1989.

[35] R. S. Sutton, "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Proceedings of the seventh international conference (1990) on machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 216–224.

[36] ——, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM SIGART Bulletin*, vol. 2, no. 4, pp. 160–163, July 1991.

[37] C. M. Gillan, S. Morein-Zamir, G. P. Urcelay, A. Sule, V. Voon, A. M. Apergis-Schoute, N. A. Fineberg, B. J. Sahakian, and T. W. Robbins, "Enhanced avoidance habits in obsessive–compulsive disorder," *Biological Psychiatry*, vol. 75, no. 8, pp. 631–638, April 2014.

[38] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[39] T. W. Robbins, C. M. Gillan, D. G. Smith, S. de Wit, and K. D. Ersche, "Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry," *Trends in cognitive sciences*, vol. 16, no. 1, pp. 81–91, 2012.