

## IJCNN2005 Talk

### TITLE

Today I will present an integrate-and-fire model of prefrontal cortical function that implements a biological mechanism of reinforcement learning. In particular, a plausible mechanism of action selection is demonstrated, as well as the conditions for which previously learned behavior is reused in a manner that can provide short-cuts in subsequent tasks.

### PROBLEM

As Miller and Cohen describe in their review: PFC and specifically orbitofrontal cortex is involved in goal directed behavior where stimuli must be associated with reward. The biological mechanisms that enable context/stimulus dependent change in action selection in prefrontal cortex are not well understood.

### HYPOTHESIS

We propose that goal-directed behavior is guided by networks of minicolumns - those are clusters of densely interconnected neurons in prefrontal cortex that retrieve known associations between specific states and actions. Our proposal amounts to a neural implementation of reinforcement learning, as described by Sutton and Barto. In this model, action selection is achieved by a convergence of activity in the minicolumns that is elicited by two sources, namely the desired goal and the current state. Because of this, short-cuts can be employed in novel learning, when provided by the spreading activity through known associations.

### TASK

- The spiking neuron model I present here is focused on simulating the behavior and activity of orbitofrontal neurons described in an experiment by Schultz, Tremblay and Hollerman.
- In this task, primates learned to give Go versus NoGo responses to randomly presented visual cues.

- In each trial, one visual cue was associated with rewarded movement (a key press resulted in a reward), a second cue with rewarded non-movement and a third with unrewarded movement. In the case of unrewarded movement, no reward is given, but only the correct response, namely Go action leads to a subsequent rewarded trial.
- Schultz et al. conducted unit recording in orbitofrontal cortex and identified specific neuronal responses that they labeled selective for the cue phase, the phase prior to reward and the reward phase.

### EXAMPLE

This is a schematic representation of one such set of trials. Each presented cue is chosen randomly and in rewarded movement and rewarded non-movement, a correct response is rewarded.

The unrewarded movement trial differs in that no immediate reward is given for a key press. Instead, a key press insures that the next trial will present a rewarded movement or rewarded non-movement cue, while another unrewarded trial will otherwise follow.

### USE

We used our model to simulate the peri-stimulus time histograms computed for spiking data obtained by Schultz et al.

### CATACOMB

Our model consists of a hierarchical network of integrate-and-fire neurons with dual-exponential synaptic responses, implemented in CATACOMB.

The environment allows us to rapidly prototype network models, while retaining the ability to specify detailed parameters such as the time course of after-hyperpolarization in a specific type of neuron.

### SPIKING

Here you see a typical example of the spiking membrane responses obtained with this modeling environment.

## DESIGN

Our model is “design-based”, as described in detail in Robert Cannon’s paper in Neuroinformatics.

We simulate (1) Actions in the environment, (2) Stimuli generated by perception, (3) Neuron dynamics in prefrontal cortex.

## PRESUPPOSITIONS

We presuppose dense connectivity within a minicolumn and sparse, possibly long distance connectivity between minicolumns. This is based on anatomical evidence.

We presuppose that activity in prefrontal cortex relates to learning of goal-directed behavior, specifically that it involves representations of past and current stimuli and of proprioceptive sensation and prediction of motor action.

We presuppose that connections within and between minicolumns can be strengthened rapidly by long-term potentiation.

## MINICOLUMNS

Here you see spreading activity from a goal representation to associated representations. When the spread reaches its end, the animation loops to the start.

- This is a network of 6 minicolumns, each of which has the same structure.
- The light blue minicolumns represent states, the yellow ones actions.
- After behaviors are learned, strengthened synaptic connections associate states and actions, representing known transitions from one state to another due to an action – as in reinforcement learning.
- Separate populations of neurons are involved in forward and backward associations between the same state and action.
- Goal-directed behavior is guided by a convergence of retrieval activity: 1. spread from a goal representation, 2. gated activation at current state neurons.
- Since multiple neurons are available in each input or output population, one state or action minicolumn can take part in multiple distinct behaviors.

- The animated example here shows retrieval for a rewarded move trial – the desire for reward spreads through known associations – when the spread reaches the current state, a corresponding output spike is generated that guides next action.

## CONVERGENCE - ACTION SELECTION

This convergence is shown here for an unrewarded movement trial:

- Current state leads to subthreshold state activity.
- The spread of goal activity adds input.
- A winner-take-all process causes one output to guide behavior.

## INPUT PROCESSING

Input from the Catacomb simulation of the experimental environment comes in the form of spike trains that represent the detection of a visual cue, the proprioception of motor action or inaction, and of reward received.

Here you see the input produced in 7 separate training trials.

Spiking neuron circuitry is used to preprocess the input, so that each change of the input results in a spike pair that represents the most recent state and action. The state-action spike pairs become the input to the network of minicolumns in our model of prefrontal cortex.

## ENCODING

Here you see the encoding of the associations between minicolumns as an unrewarded movement and then a rewarded movement trial are learned. The process would take too long to explain in detail here, but you can see the associations for the spread from the goal being strengthened as the sequence progresses.

(Or put this into an appendix and say as much.)

## STDP

Input spike pairs can appear at any time and are generally separated by intervals of at least hundreds of milliseconds.

Yet, at this plot shows, synaptic strengthening due to spike timing dependent plasticity occurs only if pre- and postsynaptic spikes occur within less than 40 ms of each other.

A spike-timing dependent potentiation rule developed in our lab was presented here yesterday by Anatoli Gorchechnikov, but in the model presented in this talk, we use a simpler implementation through a lookup table.

Also, the sequence of pre- and postsynaptic spikes needs to be repeated for sufficient strengthening.

### ADP

We propose a persistent firing buffer that compresses the interval between spikes, based on an intrinsic property – afterdepolarization – shown by Andrade in prefrontal cortex and by Klink and Alonso in entorhinal cortex, as well as model by Eric Fransen.

A spike causes afterhyperpolarization that is followed by an afterdepolarization, which can cause repeated spiking when the cell is sufficiently depolarized.

This slide shows the acetylcholine dependence of persistent firing based on afterdepolarization, as demonstrated by Klink and Alonso. Spiking does not exceed the period of stimulation with low levels of acetylcholine. Persistent spiking is maintained with high levels of acetylcholine. And in the presence of the theta brain rhythm, the persistent repetition of spikes becomes synchronized with theta rhythm.

### STM

We assume that buffer cells experience cycles of theta modulation causing relative hyperpolarization in which spike reactivation by afterdepolarization is suppressed, but spike due to novel input may appear. A subsequent phase of depolarization enables repetition of a buffered spike by afterdepolarization.

If the cue stimulus of an unrewarded movement trial is maintained in the buffer and an action spike appears as new input, then both are maintained in sequence, separated by recurrent inhibition.

The repeated spikes appear within 40 ms of each other so that buffer output can enable strengthening by spike timing dependent plasticity at synapses within and between minicolumns.

This buffer model resembles a model of short-term memory proposed by Lisman and Idiart.

Buffering based on afterdepolarization initiated at different times in some ways resembles the competitive queuing working memory presented yesterday by Dan Bullock, in that the strength in terms of membrane potential of consecutive items held in the buffer at any given time reflects their temporal order.

## FIFO

While one population of interneurons ensures the separation of buffered states and actions, a second population of interneurons achieves the suppression of first item repetition when a full buffer receives new input.

We are in the process of submitting a manuscript that describes in more detail the proposed first-in-first-out mechanism of a short-term memory buffer, as well as its sensitivity to noise and parameter changes.

## PERFORMANCE

Our model, like the mackaque monkeys, learned to perform the task correctly. This is shown in this slide for a sequence of 6 test trials. In 4 trials, motor action that leads to a key press is correctly elicited.

## RESULTS

We show comparable results for the activity recorded in 3 different orbitofrontal neurons by Schultz, Tremblay and Hollerman and the simulated activity in 3 neurons of our prefrontal model.

In this peristimulus time histogram, each row of plots presents data from one specific neuron. Within each plot, spikes are represented by black dots and each row of dots is a separate trial recorded from the same neuron.

The simulated trials contain no random variation at this time.

Plot A shows a neuron that spikes preferably in the cue or instruction phase of rewarded movement trials. Similarly, specific neurons in the simulation spike predominantly in the instruction phase of a rewarded movement trial.

Plot B shows a neuron that spikes with the highest frequency in the phase preceding reward during a rewarded movement trial. There is also some spiking in unrewarded movement trials. Similarly, specific neurons of the minicolumn simulation have the highest spike rates during that phase of a rewarded movement trial, predicting the movement action.

Plot C shows a neuron that spikes when reward is received. A similar spike plot is shown for the simulation in plot F.

### REUSE EXAMPLES

The following prediction may be used to test the validity of our model of networks of prefrontal minicolumns, especially its reliance on separate populations of input neurons and output neurons.

We predict that the manner in which a minicolumn represents an action or state that may be associated with other states and actions in multiple different task configurations can lead to the discovery of short-cuts in specific cases and not in others.

The prediction is shown in the following spatial navigation task examples:

In (A), a path from a starting location to goal G1 is learned. Then the direct route is blocked and a longer path from the starting location to goal G2 is learned. We predict that when the block is removed, the shorter path to G2 is immediately known and usable.

In (B), paths to goals G1 and G2 are learned from different starting locations and the paths meet in a single location – the crossing. If goal G2 is subsequently desired, but the starting location at the bottom of the maze is used then no path to G2 is known without additional learning, despite knowledge of all the parts of the path and the crossing.

### LEARNING IN EXAMPLES

These plots show the state-action spike pairs elicited for training in case A, and for subsequent performance of the task, as well as the buffer activity that allows the minicolumns to encode associations.

### CASE A ENVIRONMENT

The movie on the left shows the simulated rat as it learns the path to goal G1, then the path to goal G2. Finally, the simulated rat finds its way to goal G2 via the short-cut.

The movie on the right shows the development of place field activity as the virtual rat learns the maze. Different colors indicate activity of different place cells. In the present simulations, place fields are non-overlapping.

### CASE A MINICOLUMNNS

Here you see the prefrontal minicolumns involved in simulations of case A. Arrows indicate the strengthened connections that encode associations learned for paths to goals G1 and G2. The dark circles indicate neurons activated for the short-cut. Note that the short-cut uses a subset of the neurons and associations learned for the path to goal G1.

### CASE B MINICOLUMNNS

In case B, although a single minicolumn is used to represent the crossing, different input and output neurons participate in the paths to goal G1 and goal G2. At no point is an association made that could link parts of the path from the bottom of the maze to the crossing and from the crossing to goal G2.

This is a prediction specific to our proposed model of prefrontal minicolumns involved in action selection during goal directed behavior.

### CONCLUSIONS