A framework for implicit planning
Towards a cognitive/computational neuroscience theory of prefrontal cortex function

Nicola De Pisapia

PhD
Institute for Adaptive and Neural Computation
The University of Edinburgh
2004
DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.
Acknowledgement

First of all I wish to thank my main supervisor, Nigel Goddard, who has provided all the encouragement, help, support, proofreading, direction that I could wish for. My second supervisor, and also director of the Institute for Adaptive Computation when I worked on this thesis, David Willshaw, has also provided excellent support, guidance, and, most notably, has allowed me to follow my own interests since my first day in the Institute.

My thanks go to all the other members of the Institute, and in particular, in random order, to Stephen Eglen, Chris Williams, David Sterratt, David Barber, Richard Shillcock, Martin Mayer, Kit Longden, Andrew Gillies and Michalis Titsias with which I have interacted more, with plenty of instructive and entertaining discussions on topics to various extents related to this thesis.

And finally, my deepest thanks to my wife Giusy for sharing with me, as ever, both enthusiasm and sacrifices.
Abstract

In this thesis I review available experimental findings on rodents, monkeys and humans, and I suggest a unifying framework for the role of the PFC in behavior. I implement computational models to test this interpretation, and propose novel experiments.

My suggestion is, as also other researchers have proposed, that the PFC is involved in planning, i.e. in evaluating which course of actions to execute in order to reach a goal. Unlike previous researchers, I emphasize and limit the study to implicit aspects of planning, and describe a view of this process that is quite close to Instrumental Conditioning, which doesn't involve language, external measures of time (clocks), instructions or social interactions of any kind. Nonetheless implicit planning can be a quite complex activity. Under this interpretation, I show reward based computational models that, while mimicking some of the known neural properties of the PFC, can perform planning.

One aspect on which I focus is the capacity of neurons in dorsolateral PFC to code temporal information, namely when to expect task related events to occur. This is a core requirement to organize and plan complex behavior. Another aspect on which I focus is the fundamental role played in the planning process by the basal ganglia. As a plan is executed successfully several times, the basal ganglia build a chunked representations of the whole course of actions needed to reach a goal. At the same time the posterior cortex and the cerebellum retain the detailed information of how and where to execute these actions. This process allows the PFC to plan about more and more complex goals, because each action chunk, that constitutes a unit in what is
evaluated, stands for longer and longer sequences of actions. I show reward based computational models capable of producing this action chunking process.
Contents

Chapter 1 - Introduction ................................................................. 1

1.1 The general approach ............................................................. 3
1.2 Thesis summary ......................................................................... 6

Chapter 2 - The biology of the prefrontal cortex ................................. 15

2.1 Anatomy of the prefrontal cortex ............................................... 15
2.2 Ontogenesis and phylogensis of the prefrontal cortex .................. 18
2.3 Prefrontal cortex cells, connections and neurotransmitters ........... 19

Chapter 3 - Experiments and theories on prefrontal cortex role ............ 25

3.1 Lesions and psychiatric studies .................................................. 27
3.2 Cognitive tests on prefrontal cortex .......................................... 28
3.3 Imaging techniques .................................................................... 33
3.4 Cellular recordings .................................................................... 34
3.5 Computational models of the prefrontal cortex ............................ 36
3.6 Towards a unified view of prefrontal cortex function .................... 42
Chapter 4 - A framework for neural planning ........................................ 47

4.1 Hypothesis 0 - Unconscious and conscious planning ................. 48
4.2 Hypothesis 1 - Planning and action chunks .......................... 49
4.3 Hypothesis 2 - Planning and internal model formation ............ 49
4.4 Hypothesis 3 - Planning and internal model storage ............. 50
4.5 Computational framework for neural planning: modules ......... 50

Chapter 5 - Hypothesis 0: unconscious planning ............................ 53

5.1 Motivation ............................................................................ 53
5.2 Experiments on unconscious learning and decision making .... 63
5.3 Further questions on unconscious planning ......................... 68

Chapter 6 - Hypothesis 1: planning and action chunks ......................... 71

6.1 Cybernetics and control ..................................................... 73
6.2 Artificial intelligence and control ........................................ 76
6.3 Reinforcement Learning and planning ................................ 79
   6.3.1 The RL framework .................................................. 79
   6.3.2 Temporal Difference learning .................................. 81
   6.3.3 Policy iteration ...................................................... 82
   6.3.4 RL and planning ..................................................... 83
   6.3.5 Temporal abstractions and options ............................ 84
   6.3.6 Learning and options ............................................. 87
   6.3.7 Synchronous Value Iteration .................................. 89
6.4 RL simulation using temporal abstraction: a motor task .......... 90
6.5 The biology of action chunking ........................................... 93
Chapter 7 – Hypothesis 2: Planning and Internal Model Formation

7.1 Dopamine and reward based neural computations .......................... 107
7.2 A simulation of the dopaminergic role in the basal ganglia .......... 111
7.3 Contiguity versus contingency ................................................. 122
7.4 Internal models in the brain ......................................................... 129
7.5 Internal model learning ............................................................ 132
7.6 Simulation of temporal aspects of Watanabe 1996 experiments .... 138

Chapter 8 – Conclusion: Discussion and Future Directions .............. 147

8.1 A proposed experiment ............................................................. 150
8.2 Future directions ................................................................. 152

References .............................................................................. 156
Chapter 1

Introduction

The aim of my research is to investigate prefrontal cortex (PFC) activity and its role in the organization of actions in mammals. The behaviour of complex living systems is based on past experiences, but fundamentally oriented towards the future, and towards the capacity to anticipate incoming events in order to survive and to sustain the delicate equilibria of their bodies in uncertain environments.

The PFC constitutes the highest product that evolution has achieved to win the struggle to survive, and I interpret its activity as planning, that is to say a hierarchical process that controls, by means of simulations and evaluations, the order in which a sequence of goal directed actions should be executed.

My long term research goal, that goes beyond the present work, is the construction of a full and biophysically realistic computational theory of planning. This goal is highly ambitious, and it has to deal with several disciplines, including Neuroscience, Computer Science, Psychology and Philosophy. This thesis is the first step towards that goal.

**************

In this thesis I review available behavioural experiments on rodents, monkeys and humans involving PFC activity. I suggest a unifying interpretation of the role of the
PFC in behaviour, and I implement computational models to test it. I also propose some experiments to further test the interpretation, and novel theoretical possibilities. The view that emerges from the present work is that the PFC is involved in planning, i.e. in generating, evaluating and selecting which course of actions possibly to execute in order to reach a goal. With specific regard to humans, unlike most of the other researchers in cognitive neuroscience, I emphasize the necessity to study mainly *implicit* aspects of planning, which don't necessarily involve language, external measures of time (like clocks or calendars), instructions or social interactions of any kind. Throughout this thesis I use the term implicit interchangeably with *unconscious*, when I want to underline an opposition to conscious aspects.

Traditional studies focus on conscious and mostly language based organization of complex actions, whereas I pursue the idea that, in order to understand these high level cognitive and social capacities to plan, it is necessary to understand implicit aspects first. Therefore, the term planning in the present work denotes not only a conscious capacity, but mainly an unconscious one. In this thesis, I restrict the analysis to the unconscious planning capacity, and leave the conscious aspects for future work.

I show reward based computational models that, while mimicking some of the known neural properties of the mammalian PFC, can perform implicit planning. One modelling aspect on which I focus is the capacity of neurons in Dorsolateral PFC to code temporal information, namely *when* to expect task related events to occur at different time scales, thus enabling Temporal Abstraction (Precup 2000). This is a core computational requirement to organize and plan complex behaviour. Another modelling aspect on which I focus is the fundamental role played in the planning process by the basal ganglia. As a plan is executed successfully several times, the basal ganglia build a chunked representations of the whole course of actions needed to reach a goal (Graybiel 1998). At the same time the posterior cortex, i.e. non prefrontal cortex, and the cerebellum retain the detailed information of the execution of these actions. This process allows the PFC to plan about more and more complex goals, because each action chunk, that constitutes a unit in what is evaluated, stands...
for longer and longer sequences of actions. I show reward based computational models capable of producing this action chunking process.

Similar ideas were exposed in the influential *Plans and the structure of behaviour* by Miller, Galanter and Pribram (1960), but that book had two fundamental limitations: one is the poverty of experimental evidence to justify the view of the PFC role (the references in the book are all from the 50s backwards, and therefore not many experimental techniques to study the brain were available at the time), and the other is that the formal mathematical aspects are mostly omitted, and only general ideas are illustrated. The ideas I discuss in this thesis can be seen as an attempt to overcome these two limitations, the former by putting together and reinterpreting recent experimental support for such a view of planning, the latter by proposing a more detailed framework, adopting specific mathematical techniques and constructing computational models.

At the end of the thesis I will have answered some questions and raised several new and interesting ones. In the course of this, several ideas will have been illustrated to clarify where and how to look for answers.

### 1.1 General approach

The aim of this thesis is to design computational models of PFC activity, both at the neural and cognitive level. This aim lies within the general approach called computational cognitive neuroscience by some (O'Reilly 2000, for example), or neurobiological modelling by others (Montague and Dayan 1998).

This approach has been developed within the last twenty years. The basic idea is that cognitive models and neuroscientific models of specific brain regions can be combined into a single computational model. That is, neurocomputational models can be the result of an integrated description of both cognitive and biological aspects of neural activity.

The two basic ideas behind this approach are:
(1) With regard to its role in cognition and behaviour, the activity of single neurons can be interpreted as that of simple processors of information (in the form of spike rates). By extension, the role of networks of neurons in cognition is that of a complex parallel computing machine. If a monkey has to learn to press a left button 5 seconds after a red light is showed, and to press a right button after a green light is showed, that can be modelled as a computational capacity of the animal’s neurons to code appropriate actions and time delays. Essentially, in this approach, computation is the link between the mind and the brain. This idea is quite old, it was formulated even before computers were actually available (in G. Leibnitz, for example), but it has been elaborated only under the influence of Cognitive Science (see Scheutz 2002 for a recent discussion).

(2) Computational models can be used and combined to describe how different neural tissues work, both at the cognitive level (top-down) and at the biological level (bottom-up) (Montague and Dayan 1998).

Both ideas have been highly debated in the last few decades, and a full discussion of them is outside the scope of this thesis, but I do want to report a few comments on the second one. The idea that cognitive and neurobiological models should go together is still in its infancy, and some scientists criticize it by saying essentially that, given that the cognitive models and the biological models per se are not well established, to put ideas from these approaches together is inevitably even more controversial. The position in computational cognitive neuroscience is the reverse, in the sense that the attempt to find a common language for cognitive and neural description is seen as a necessity to understand how the brain works. Puzzling findings for cognitive scientists can be explained in neurobiological terms, and vice versa, neural activity in itself cannot fully explain cognitive functions. The first to explicitly formulate the necessity for such a joint approach to understand the brain was D. Marr (Marr 1982). In particular he applied this to the study of vision, but he opened the possibility to extend his reasoning also to the study of any cognitive process. He reports how, until the 70s, scientists (e.g., H. Barlow) were so excited by
the successful findings in neurobiology, that they were wrongly led to believe that
descriptions of even single nerve activities and reciprocal influences were all that was
needed to understand the functions of the nervous system.

“There is nothing else looking at or controlling this activity, which must therefore
provide a basis for understanding how the brain controls behaviour” (Barlow 1972).

But in the 80s, the assertion that “there is nothing else” was challenged by the
difficulty to find convincing explanations for the way in which maps of neurons work.
“Something else” appeared to be needed, that is to say an higher level of explanation
that includes the general goal of the computation that is taking place in the map of
neurons.

Marr, to better illustrate this point, introduced three (quite famous) levels at which the
brain, interpreted as an information processing device, can be understood. At the top
level there is the abstract theory of what is the goal of the computation taking place in
that nerve tissue, and what is the general strategy to achieve it. This level is
essentially the cognitive description of the function under analysis. The middle level
is the representational choice, i.e., the specific algorithm used to achieve the goal
described at the computational level (e.g., various artificial neural networks, Turing
machines, Gandy machines or any abstract computing device). Finally, there is the
lower level, that specifies the actual physical implementation of the specific algorithm
(chosen in the previous level), which can be an artefact or a biological system.
Computer scientists are usually more attracted to the study of the top and middle level
(as Marr did, for example), whereas neurobiologists and engineers are inevitably more
focussed on the lower. But a computational cognitive neuroscientist has to deal with
all of them.

The overall approach of this thesis is, indeed, to investigate PFC properties at the
neural level, and separately at the cognitive level, with the aim of unifying them at the
computational level. Most of the work can be described as top-down, but several
bottom-up considerations will be key. In particular, the questions I address are:
Introduction

- Which are basic biological features of the PFC?
- What is the cognitive role of the PFC?
- What are key computational solutions that any computing device performing this role must adopt?
- Can the PFC implement these solutions?

The cognitive role I ascribe to PFC is planning (as I mentioned, not necessarily intended as a conscious process, or under the directions of conscious will). Computational requirements of simple (non symbolic) planning devices are discussed, and in particular I focus on a Reinforcement Learning (Sutton and Barto 1998) approach to planning. Specific algorithms that I investigate to implement PFC activity and dopaminergic signals use Temporal Difference Learning and Temporal Abstraction (Precup 2000). These algorithms constitute the result of constraining the cognitive and biological requirements together in working computational models.

1.2 Thesis summary

In the rest of this introduction I present a general summary of the thesis, chapter by chapter.

In the second chapter, using selected sources, I briefly review basic biological aspects of the PFC: elements of its anatomy, and comparative anatomy (which are the analogies between different species), its connections to other parts of the brain, and the main neurotransmitters involved. The PFC is the most anterior part of the brain, covering one sixth of the entire cortical area. It has feedback connections, including some contralateral via the corpus callosum, with many cortical and sub-cortical areas, particularly with the premotor and motor areas, temporal and parietal cortices, thalamus, hippocampus, basal ganglia, amygdala, cerebellum, and brainstem. The
main neurotransmitters found in the PFC are dopamine, serotonin, acetilcoline. In the first chapter of the thesis we see these biological aspects in more details.

In the third chapter I review several experimental traditions that have put different emphasis on what might be the role of the PFC in behaviour. The main investigative methods in the literature have been:

- Cellular recordings *in vivo* in monkeys.
- Lesion studies in monkeys and humans.
- Psychiatric studies (mostly on schizophrenics).
- Neuroimaging (fMRI, PET).
- Drugs that act on the PFC.
- Cognitive tests.

With regard to the last method, throughout the years researchers have conceived a number of *ad hoc* cognitive tests, believed to engage mainly the PFC. Therefore subjects with lesions in that area were expected to show impairments in the performance of these tests if compared to healthy individuals. I summarize these tests, and indicate the main psychological processes believed to be engaged.

I review the ideas of P. Goldman-Rakic and her working memory interpretation, J. Fuster and his Action-Perception cycle, E. Miller and the *what* and *where* distinct processing of information, A. M. Owen and conscious planning, J. Cohen and active representations, S. Dehane and J. P. Changeux and the hierarchical organization, T. Shallice and the temporal organization of behaviour. In this chapter I also review and critique the main connectionist models on PFC activity, and what they can achieve.
In the fourth chapter I integrate and re-interpret the reviewed experimental and theoretical studies on the functions of the PFC. My view is that the role of the PFC is planning, namely the process of producing, evaluating and selecting the best course of actions that has to be taken in order to achieve a goal. As mentioned, I use the word *planning*, which mostly in cognitive psychology refers to a conscious process, as denoting also, and in the present work mainly, an unconscious one.

This chapter introduces the architecture of the computational model I propose. I maintain that, in order to account for the more complex role played by the PFC in the organization of behaviour, a generalization of the simple activity of the basal ganglia as previously modelled in Montague et al. (1996) is needed. I propose that this can be theoretically given by a very recent mathematical extension of more basic Reinforcement Learning techniques (including Temporal Difference learning) (Sutton and Barto 1998), called Temporal Abstraction (Precup 2000).

The PFC is traditionally tripartite into dorsolateral, ventromedial and orbitofrontal sub-parts. I interpret the dorsolateral PFC as the part of the PFC which is mainly involved in sustaining through time the representations of information stored in posterior regions of the cortex and cerebellum, implementing in this way the capacity to keep information active, and to manipulate it in order to achieve goals. This is accomplished mainly by constructing on-line internal models of the world, namely models of how the environment changes through time according to which actions are taken.

Working memory and on line construction of new Internal models of the environment are the same process, with the second being a more precise and implementable definition. I interpret the ventromedial PFC as mainly involved in emotional evaluations while processing goal-oriented behaviour. The orbitofrontal cortex is instead interpreted as central in socially modulated behaviour, and in individual evaluations of relative preferences between stimuli. Feedback signals (errors in the predictions of both rewards and outcomes) come from dopamine, and they are the basis for the learning process.

These different aspects of planning can be integrated into a unified view of the PFC as the part of the brain of highly complex animals, where control of the overall goal
directed behaviour emerges as a coordination, in uncertain environments and in relationship with other members of the same species, between rich lower level sensory inputs and motivations.

In this interpretation, we also need to review other areas of the brain, strongly connected to the PFC, which inevitably have a role in the planning process. In particular I discuss the basal ganglia, which have a major role in chunking actions, thus constructing behavioural repertoires, and the posterior cortex and cerebellum, that consolidates internal models of the world once the activation based models in the PFC have been several times successful in their predictions.

The idea of chunking in the basal ganglia has been suggested by A. M. Graybiel, and tested in interesting works in the striatum. I will discuss this work at length in the thesis.

The idea that there is a shift of the neural activity from frontal areas of the brain to posterior areas has been experimentally tested on motor tasks by R. Shadmehr. He used PET (Positron Emission Tomography) to scan subjects in different phases during the execution of a manual task (they had to move the handle of a robot manipulandum towards target locations and against opposing mechanical forces). While performance remained unchanged, the brain engaged different areas during different phases: at the beginning of the learning process PFC regions were more active, whereas 6 hours after practice there was a shift to the premotor, posterior parietal and cerebellar structures. The authors suggest that this shift might indicate a change in the neural representation of the internal models and an increase in the functional stability.

Notice here that this task is fundamentally motor. Planning fine grain movements of the body does not probably involve PFC activity only, but, being such movements an important activity for survival, a specialized area on its own, namely the cerebellum, is also centrally involved.

An hypothesis that is quite legitimate to postulate, even though it is not explored in this thesis, is that the main difference in the roles of the cerebellum and the PFC is based on timing issues. While the PFC would be devoted to the construction of internal models in the long time scale (in the order at least of a few seconds), which is indeed discussed in this work, the cerebellum might be involved in the construction
and storage of internal models in the short time scale (in the order of milliseconds), thus involving mainly fine grain movements and control of muscles. With regard to the timing issues and the cerebellum there is abundant literature (as in Ivry et al. 1989), and also neurocomputational modelling (Kawato 1999), in particular on internal models construction and storage in the cerebellum for motor control and trajectory planning.

***************

I organize the general architecture for PFC activity and planning into four basic hypotheses:

*Hypothesis 0 – Implicit planning*

In humans, besides the well known conscious planning capacity, based on communication with co-specifics, external measures of time and internal speech, there is a basic unconscious and implicit adaptive capacity of their brains to simulate and evaluate goal oriented behaviour and consequent experience, in order to possibly execute the best (given current knowledge) sequence of actions, that should be performed in order to achieve a goal.

*Hypothesis 1 – Planning and action chunks*

This planning capacity can be modelled as a processing of information, under the directions of reward seeking behaviour, and with the fundamental capacity to chunk successful sequences of goal oriented actions. This chunking process facilitates and speeds up the planning process, because these chunked sequences of actions don’t need to be re-evaluated, but are well acquired knowledge about how to go from one state of the environment (start) to another (end/goal/reward).

*Hypothesis 2 – Planning and internal model formation*
Planning is achieved through phasic dopaminergic effects, acting on the PFC as an error signal in the prediction of rewards and of every task related stimulus, thus allowing the construction of the internal models. Novel internal models of task relevant information on the environment are thus constructed and maintained in the Dorsolateral PFC in the form of activation based processes. In these internal models information about where, what and when goal related events are going to happen is learned and put together.

**Hypothesis 3 – Planning and internal model storage**

Lastly, posterior cortex (namely, every part of the neocortex which is not PFC) and cerebellum consolidate internal model representations of the world once the activation based models in the PFC have been several times successful in their predictions. This is a key computational passage that allows efficient planning. It is the essential capacity to store knowledge of execution of well acquired (sub)tasks, so that, when evaluating new courses of actions, the steps and the state transitions stored do not need to be re-computed, but only information about the start state and the end state is used.

The first three hypotheses are discussed at length in different subsequent chapters, while the last one is only briefly discussed in chapter 4. Several computer simulations relating to these hypotheses are constructed and described at different points of the thesis when relevant.

In the fifth chapter, a rather philosophical one, I discuss hypothesis 0. I define better the cognitive planning capacity, distinguishing clearly between conscious and unconscious aspects. I discuss how cognitive neuroscience research on human PFC activity should free itself from a prejudice, namely that planning, with its association with decisions and will, must be exclusively studied as a conscious activity, explicit and communicable. Whereas unconscious perception is well studied (as in blindsight
for example), unconscious planning seems an oxymoron to most researchers. I argument that the essential components that make humans organize goal-oriented courses of actions are not necessarily consciously accessible, not just in activities in which we are experts (and which do not need monitoring), but in every situation. Mostly humans don't know much about the knowledge that they instead have on how to achieve goals, and also they might not be aware of what their goal is, even if they are constantly trying to pursue its rewarding value.

Perhaps the closest term in psychology to implicit planning is instrumental conditioning, but the highly experimental aims of researchers working in this field and the exclusive application to animals has hidden the relationship to human planning capacities.

Conscious planning is a complex activity, probably based on the unconscious one, but involving also several social and cultural aspects, like language, expectations from other individuals, the use of tools to take note of goals (pen and paper), and also of tools to measure or label time (clocks or calendars). Behind this complex activity there is a simpler, more fundamental, unconscious and individual planning capacity of the brain, which involves the construction of short term expectations about when, where, and what events will occur in the future depending on which actions are executed (the knowledge contained in the internal models). Perhaps also long term expectations can be unconscious, but this might be harder to study for the enormous number of intervening factors.

Implicit planning involves:

- the organization of these expectations in hierarchies;

- the modulation of relevance of goals depending on the current state of the body (motivation);

- the capacity to perform all these activities at the same time.
I show some of the few experiments that support or start to explore this unconscious and conscious distinction in planning, and discuss how a study of the basic neural processes involved in the unconscious planning helps in a better understanding of the conscious one.

In the sixth chapter I discuss hypothesis 1, and review some concepts about planning from a computational point of view. In general, a plan consists of two parts: (the representation of) the sequence of actions that must be taken to achieve a goal, and an internal model of how the environment changes when these actions are taken. Planning refers to the capacity of an agent (natural or artificial) to produce and evaluate a number of possible plans.

I review ideas from cybernetics (mostly about the role of feedback in reaching a goal state), from artificial intelligence (there is an abundance of literature on this topic, because planning has been a core issue of this approach), and finally from cognitive science (mainly a fundamental study by Miller, Galanter and Pribram 1960, and Newell and Simon's work on problem solving). Then I introduce a specific approach to artificial intelligence called Reinforcement Learning (RL) (Sutton and Barto 1998), and how this approach can give some interesting indications of how to implement a simple neural system for planning. I describe the main intuition and formalism behind a new development in RL, Temporal Abstraction (Precup 2000).

It is interesting that some previous versions of RL (Montague et al. 1996) have already been successful in modelling the basic interaction between the basal ganglia and the frontal areas. Therefore, we have several clues that motivate us to investigate the use of these new RL ideas, and to see how much they can help in the understanding of the computational requirements for planning in the PFC. This chapter includes a brief history of the idea of an internal model, psychological studies from K. Craick (1943), and some control theory.

In the seventh chapter I discuss hypothesis 2. I review experiments and models in animal conditioning using rewards, which theories of the PFC should take into account. Ideas from Hull, Tolman and Skinner are discussed. Animal learning studies,
in particular by Balleine and Dickinson (1998), show that Skinnerian stimulus-response habit mechanisms, producing the so-called superstitious behaviour (Skinner 1948), cannot fully explain goal directed behaviour as demonstrated in mammals. In particular, one fundamental missing aspect is the capacity of mammals to learn the contingencies (causal relationships) between stimuli and actions. They must know and evaluate the consequent states of the environment depending on their actions before they actually execute them. Stimulus-response habits can instead only detect temporal contiguity. The capacity to "understand" the causal relationship presupposes the use of an internal model of the world that predicts the temporal evolution of the environment. Lesion studies in the dorsolateral PFC of rats suggest that this is the part of the brain mainly involved in contingency learning. I explore the ideas of this chapter with several simulations.

Finally, in the eighth chapter I discuss the main contributions of this dissertation to the understanding of PFC activity, a novel experiment to test temporal expectations of not directly rewarding stimuli, and future directions of research.
Chapter 2

The biology of the prefrontal cortex

Prompted by sensory-associational and motivational inputs of cortical, subcortical, and limbic origin, the prefrontal cortex critically intervenes in the initiation and guidance of the behavioral structure. At the start of the action and in the course of it, some of its representational cell networks become operational networks. As such, they guide action to its completion. Through the interactions of these networks with neuronal populations in other cortical and subcortical regions, the prefrontal cortex ensures the retention of information for prospective acts, prepares neural systems for those acts, and suppresses extraneous influences that could lead the action astray.

Joaquin M. Fuster. The Prefrontal Cortex (1989)

In this chapter I review basic biological facts about the PFC, including the connections of the frontal lobes with other areas of the brain, and some interspecies comparative aspects.

2.1 Anatomy of the PFC

The main focus of this chapter and the following ones is on the brain of humans, but I consider also data on several other mammals, in particular rodents, and other primates. This use of data from animals is a common method in science concerning humans, given the limited possibility of intervention due to obvious ethical constraints. But we should not forget that interspecies differences are substantial, and some might be still unknown. Definitions of differences in the structures of several
nervous systems, from high brain regions to neuronal structures, such as neurotransmitters or ionic channels, are still a matter of ongoing research. But it is clear enough that the morphology of the PFC varies considerably from species to species, and therefore only rough homologies can be established.

What is the PFC? Although it is highly debated (Uylings et al. 1990), the oldest and most common anatomical definition of PFC in all species is:

*the part of the cortex that receives projections from the mediodorsal nucleus of the thalamus* (Rose et al. 1948).

It is a subregion of the most anterior lobes of the cortex, namely the frontal lobes (yellow in Fig. 1 and 2).

In humans, the border of the frontal lobes with the parietal lobes (green in Fig. 1 and 2) is called the *central sulcus*, and with the temporal lobes (red in Fig. 1 and 2) is called the *lateral fissure*. The frontal lobes, besides the PFC, also include two other
sub-regions: the motor cortex (area 4), and the secondary motor zones, which include premotor and supplementary motor cortex (area 6), the frontal eye field (area 8), Broca's area (area 44 and 45), and the posterior cingulate cortex (area 24, in Fig. 2). The remaining areas of the frontal lobes constitute the PFC. The PFC occupies one half of the frontal lobes, which themselves constitute one third of the entire cortex.

The PFC is traditionally divided into three subregions (roughly represented in Fig 3, next page) (Cummings 1995), called the dorsolateral PFC (green), the orbitofrontal (sometimes called limbic) PFC (yellow), and the ventral PFC (red). The reason for such a subdivision is that each of these three different areas is involved in a different circuit with other cortical and sub-cortical structures.

All these definitions, although quite common, must not be taken too literally, but only as general indicators, because each method has some kind of limitation. For instance, the commonly used Brodmann system was based on the cytoarchitecture of one single brain. Also, in the definitions I have just given, that are reported in most manuals, it is not clear if area 45 must be ascribed to PFC (being part of the Dorsolateral PFC) or to
the secondary motor areas. At any rate, I will adopt the definitions articulated above, but it must be kept in mind that serious work on clear and coherent labeling of brain regions according to different criteria is still in progress.

Fig. 3
Divisions of the prefrontal cortex, lateral view on the left and internal view on the right. The green area represents the dorsolateral part, the red area represents the ventral part, and the yellow area represents the orbitofrontal part.

2.2 Ontogenesis and phylogenesis of the PFC

During human evolution, the PFC has expanded more, relatively to the rest of the cortex. Comparative anatomy has traditionally stated that the human PFC is relatively of bigger size than that of any other animal, and in particular of any other hominoid, living or extinct (Deacon 1988 and Rilling 1999). This claim is generally considered valid, but recent findings (Semedeferi et al. 2000), which make use of Magnetic Resonance Imaging, show that humans have comparatively frontal lobes of the same size of the great apes (gorillas and orangutans).

This study doesn't exclude the possibility that a difference between the Frontal Lobes of the great apes and the humans may reside in a relatively bigger human PFC, but the authors claim that the absence of macroscopic landmarks delineating the borders of the Prefrontal areas makes these comparative studies extremely difficult. In any case,
humans and monkeys have externally homologous Prefrontal areas, and this helps in
the study of the human brain because of the mapping between the Brodmann areas of
the two species (Goldman-Rakic 1987).

With regard to the ontogenesis of the PFC, cytological studies show that, even though
if infants have this part of the cortex well developed (Mrzljak et al. 1990), it is only
during adolescence that the number of connections explodes (Lewis 1997). This
number settles down only during late adolescence. And finally, in the elderly, the
natural involution of the whole brain starts indeed in the PFC: this includes neuronal
death, dendritic atrophy, loss of synaptic spines.

2.3 PFC cells, connections and neurotransmitters

The cytoarchitecture of the PFC (Fuster 1989) shares with the rest of the cortex the
typical organization in six layers (Economo 1929). In motor and posterior areas of the
cortex, layer V is thick and it has big pyramidal cells, and layer IV is thin and without
granular cells; in the PFC instead, as it progresses from premotor areas towards the
frontal pole, layer V becomes thinner and without giant pyramidal cells, and layer IV
becomes thicker and primarily constituted of granule cells. In layer IV, in particular,
the predominant kind of granule cell is the stellate cell, characterized by short axons
and short dendrites, so that it branches out only in the proximity of its soma.

With regard to local connections, the PFC pyramidal cells appear to be organized in
stripes (Levitt et al. 1993), each one relatively isolated from the other. Both the high
number of stellate cells and the stripes of pyramidal cells seem to suggest strong
recurrent connections, thus supporting the view, that I also adopt, that the PFC has the
role of maintaining activations in the face of disruption due to noise or distracting
stimuli.

The connections of the PFC to cortical motor areas are not direct, but mediated
through the basal ganglia and the thalamus. The general links of the PFC to motor
areas is shown in Fig. 4 (adapted from Fuster 1989).
The PFC doesn’t receive direct inputs from the sensory periphery, but the sensory information reaching it has already been processed by other cortical or subcortical areas. Different PFC subregions specialize in different cortical sensory afferences: areas 46 and 12 for visual inputs, area 8, central and dorsal part of area 46, and area 10 for auditory inputs, ventral area 46 and area 12 for somatosensory inputs. Orbitofrontal cortex seems to be more distributed and polymodal, receiving connections from visual, auditory, somatosensory, gustatory and olfactory cortical areas (Barbas 2000).

With regard to subcortical areas, the PFC sends connections, ipsi as well as contralateral, via the corpus callosum, to the basal ganglia, subthalamus, hypothalamus (as shown in Fig. 4) and it has feedback connections with the thalamic nuclei, hippocampus, amygdala (in particular with the orbitofrontal part), cerebellum, and brainstem, including the ventral tegmental area.

In general, three main parallel circuits (Cummings 1995), involving the three different subregions of the PFC, can be detected.

The first circuit is the focus of interest in the models I develop in this thesis. This circuit, starting in the dorsolateral part of the frontal lobes also receives inputs from, and sends output to, the three other lobes of the cortex. The entire cortex, including
the dorsolateral PFC, projects unidirectionally to the caudate nucleus in the striatum, and to distinct areas of substantia nigra and the globus pallidus. The caudate projects directly, and indirectly through the subthalamic nucleus, to the medialdorsal nucleus of the thalamus, that then projects back to the dorsolateral PFC. The detailed picture, with all the afferents and efferents, of this circuit is depicted in Figure 5.

A second circuit includes the ventromedial cortex, the nucleus accumbens, distinct areas of the medial globus pallidus and substantia nigra, and the dorsal medial nuclei of the thalamus. In Figure 6 there are depicted all the afferents and efferents of these structures.
Finally the third circuit includes the orbitofrontal cortex, the ventral caudate nucleus of the thalamus, distinct areas of the medial globus pallidus and substantia nigra, and the ventral anterior and dorsal medial nuclei of the thalamus. In Figure 7 there is the detailed picture.

It is notable that the three prefrontal regions are linked with each other only at the cortical level, whereas the circuits in which they are involved do not share any structure at subcortical levels. It suggests that the dorsolateral PFC, in accord with our model, is in a strategic position to integrate information from the three circuits,
because it has connections with orbitofrontal and ventromedial cortex which are not directly connected between them.

As in any other part of the cortex, several neurotransmitters are present in the PFC (what follows is from Fuster 1989). Corticostriatal and thalamocortical connections use excitatory glutamatergic projections. Within the dorsolateral PFC circuit, the striatopallidal and pallidothalamic connections, as well as the striatum to pallidum and the globus pallidus externa to subthalamic nucleus, involve GABA.
Intrinsic acetylcholine is also present in the striatum. Back to the PFC, there are noradrenergic and serotonergic connections coming from the brainstem, but these are more abundant in other areas of the cortex, respectively in somatosensory and sensory areas. Also acetylcholine is present, especially abundant in layers III to VI, with projections coming from the basal forebrain cholinergic system. GABA is also found in the PFC, acting in wide-arbor and chandelier type interneurons, with an inhibitory role.

More distinctive of the PFC area are dopaminergic connections, more abundant than in any other areas of the cortex, coming from the ventral tegmental area, acting mainly on the dendrites of pyramidal cells in layers III and V. This neurotransmitter, as we will see later in the thesis, plays a key role in our models as an error signal in the prediction of rewards and also any task related stimuli.

The neurotransmission in the PFC develops very early in life, being present already at birth. Through infancy and adolescence, noradrenergic, serotonergic and dopaminergic actions develop more and more. In the elderly, the overall metabolic activity decreases, including also the activity of the neurotransmitters.

Besides the cited natural involution of the PFC due to aging, or damage that can occur due to accidents, several neurological diseases can involve this area at any age. For a general review of these illnesses, especially those related to the frontostriatal circuit, see the excellent review by Bradshaw 2001. The illnesses analyzed in that book are autism, depression, schizophrenia, ADHD (attention-deficit hyperactivity disorder), Tourette's Syndrome and obsessive-compulsive disorder. Degenerative diseases of PFC cells, such as Alzheimer or Huntington, start in frontal areas, and then extend to the rest of the brain.
Chapter 3

Experiments and theories on prefrontal cortex role

The function of the frontal lobes seems destined to be one of the most untreatable neuro-psychological problems.

Tim Shallice, From neuropsychology to mental structure (1988)

The cognitive activity of the PFC can be divided into three domains, according to the three circuits and subregions described in the previous chapter. The dorsolateral PFC is the part of the brain which is mainly involved in sustaining the representations of information stored in posterior regions of the cortex, implementing in this way the capacity to keep information active, and to manipulate it in order to achieve goals. This kind of activity can be considered equivalent to working memory. The ventromedial frontal cortex is instead thought to be mainly involved in emotional evaluations while processing goal-oriented behaviour. Finally the orbitofrontal cortex is believed to be central in socially modulated behaviour, and in evaluations of relative preferences. These apparently different roles can be integrated into a unified view of the PFC as the place of the brain of highly complex animals where control of the overall goal directed behaviour emerges, as a coordination, in changing environments and in relationship with other members of the same species, between rich sensory inputs and internal motivations. But exactly how this goal oriented
behaviour emerges from the activity of the neurons is still an open question, to which detailed models have not yet given an answer.

Many researchers (which I am going to mention in the next sections of this chapter) have interpreted the role of dorsolateral PFC as mainly in a working memory (WM) function. WM is the short term capacity to store mental representations, and the limited capacity of performing goal oriented operations on them. These mental representations can be either sensory inputs, or memories stored in the long-term memory, or recently learned motor activities.

This definition presents many ambiguities, and some researchers doubt it has a scientific validity. Is there a distinction between WM and short-term memory? Is it scientifically interesting to study a phenomenon that involves indirectly so many other cognitive capacities (sensory representations, motor commands, high level cognition, memories of other kind), to require likely the entire brain to be performed? What is the relationship between WM and consciousness, i.e., are people always aware of the content of their WM?

The most influential psychological model of WM in humans was presented by A. Baddeley and Hitch in 1974, later on refined and extended in A. Baddeley 1986. This model, which was based on several psychological experiments, consists of a main system, the central executive, that controls two subordinate systems, the phonological loop and the visuospatial sketchpad. The role of the central executive is to regulate the interactions between the two subordinate systems and the long-term memory. The phonological loop is an acoustic system that codes information in WM. It is composed of two parts: a short term acoustic store of sound inputs, and an articulatory component, whose function is to sub-vocally rehearsal items that must be remembered in the short-term. The visuospatial sketchpad acts as the articulatory loop, but instead of the acoustic inputs, it acts on visually based representations.

Since this early model, new theories of WM have been developed, and complex experiments have been realized (see Miyake et al. 1999 for an excellent review), but no researcher has definitively cleared the concept of WM from its many ambiguities. Besides the questions we asked before, we might ask: how can an homuncular theory of WM be avoided: namely, who controls the controlling capacity? Is WM a unitary
system, or it is a modular system? What is the role of WM in complex cognitive activities?

The first time that the term WM was introduced (by Baddeley and Hitch in 1974), it was also clearly speculated that the frontal lobes are the part of the brain to be mainly involved in this activity.

In this work I hope to show that, while WM is one important concept useful to describe some aspects of PFC function, it is nonetheless quite limited. To identify the role of dorsolateral PFC as a place for temporary storage and manipulation of representations, without telling what this representations are, is only telling part of the story, and more comprehensive concepts and definitions are required to capture the full range of observed the PFC roles.

The idea I put forward in this thesis is a much better hypothesis is to look for internal model representations as the form taken by WM contents. Internal models capture all aspects that WM captures, like the idea of sensorial and motor representations kept active and possibly manipulated, but their definition, as given in this thesis and also in the Reinforcement Learning literature, gives a much more precise description of how these stored representations are coded and how they are linked to actions, i.e. as predicted states of the environment depending on which actions are performed.

Furthermore the formal definition of internal model is linked to that of chunk, that is to say what is considered as a primitive level action for the specific internal model (each chunk is then constituted by lower level chunks). Thus the internal model representation is also a key part of the general chunking process. Very recent experimental evidence shows indeed that WM for spatial locations involves PFC if and only if a strategic chunking strategy takes place (Bor et al. 2003).

Note how this use of internal model as the way in which WM representations are actually coded is very explicit in the so called motor working memory literature, i.e. on working memory for motor learning and control (as in Shadmer et al. 1997, or Wigmore et al. 2002), but it is not common at all in studies on more traditional cognitive WM.

***************
In the rest of this chapter I review experiments on the PFC that have been used to construct a theory on its role in behaviour. There are essentially three ways in which scientists can experiment on the PFC: lesions studies, cellular recordings and cognitive tests (with or without techniques to measure brain activity, e.g., EEG or neuroimaging).

3.1 Lesions and Psychiatric Studies

Perhaps the oldest way to study the cognitive role of the PFC is to test the behaviour of patients who, for accidental reasons or as a result of voluntary brain surgery, have one or more lesions in this area of the brain.

The first case to be registered is today extremely famous (Damasio 1994), and it has become paradigmatic of PFC deficits. It is the case of Phineas Gage, a man who in 1848, while working in the construction of a railroad in Vermont, accidentally was hit by an iron bar, that entered below his left eye and exited from the top of his head, destroying a considerable amount of the medial region of his PFC. Quite surprisingly, he survived the accident, and, after a few days, he completely recovered his physical health, including his motor and perceptual functions. What changed radically was his personality. Contrary to his previous behaviour, he became highly disrespectful of social rules, and unable to make plans for future operations in his work. "A child in his intellectual capacity and manifestations, he has the animal passion of a strong man" was reported by his friends at the time. In a few words he was "no longer Gage".

Other studies of PFC lesions (Benson et al. 1975), which involve dorsolateral areas of the PFC, show that patients acquire pseudo-depressive patterns, with lack of enthusiasm and energy for any kind of activity. Another source of study comes from the data collected on thousands of patients with severe mental illnesses who, in the first part of the last century, were treated with the surgical removal of most of their frontal lobes (prefrontal leucotomy, see Partridge 1950). This medical practice was abandoned because it proved not to be effective.
PFC activity can also be studied, in a similar manner, on psychiatric patients whose illnesses are believed to involve the neurotransmission in PFC, as in Depression and Schizophrenia, or to induce a progressive degeneration of PFC cells, as in Alzheimer or Huntington disease.

Are all the cognitive deficits observed in PFC patients of a similar nature? Is there an unitary way to describe them? The most evident common finding is that, at first sight, as happened for Gage, their general intelligence, as for example measured by I.Q., appears to be close to normal. Also their motor and perceptive abilities remain in the normal range. In general, their deficits are not immediately detectable, especially in a laboratory, where the subjects are usually in a relatively simple and controlled environment. But in real life situations they experience serious difficulties.

In fact, usually they cannot hold their job. If married before the damage occurred, they typically divorce soon after it; they take hours to do very simple actions (such as cleaning themselves); they are unable to understand non-literal meaning of sentences; they don't know how to handle money; et cetera.

Specific clinical psychological tests had to be devised to detect more precisely their deficits, and we will look at some of them in this paper. In the end it will be clear that PFC patients fail in a huge number of situations.

3.2 Cognitive tests on PFC

A standard test for PFC damage is the word fluency test: subjects are given 5 minutes to write as many words as possible starting with a pre-specified letter ("s" for example), and 4 minutes to write as many four-letter words as possible starting with another letter ("e" for example). Prefrontal patients usually perform very poorly, as compared to normal subjects, mainly by having very few words, and also by making errors (that is, words starting with a non required letter).

Another common test is the stroop test. Subjects are requested first to name the colour in which words of colours written, as in RED GREEN BLUE, where colour and word are coincident. Then the subjects are requested to do the same thing with orthographic forms of colours words in which colour and word are not coincident, as
Experiments and theories on prefrontal cortex function

in **GREEN BLUE RED.** The first part of this test is usually performed quite easily, both by patients and healthy subjects. The second part is less easy, because of the competition between the more automatic tendency to read words, against the fact that the word does not refer to the colour that must be named. Normal subjects usually take a little more time to perform the second task, but they can manage, whereas PFC patients, especially those with damage to the dorsolateral PFC, perform very poorly. Reaction times are hugely prolonged, and also they make many mistakes, i.e. the automatic tendency to read the word, instead of naming the colour, prevails.

At a semantic level, it is possible to test PFC subjects with the *Hayling A and B* tests, conceived and studied by Paul Burgess and Tim Shallice. Subjects are presented with a series of sentences with the last word missing (e.g., "He mailed a letter without ..."); in Hayling A they have to complete them with a suitable word (e.g., "a stamp"), whereas in Hayling B they have to complete them with any word that makes no sense (e.g., "an elephant"). PFC patients are slower than healthy subjects on Hayling A, but on Hayling B they are excessively slow. Given a sentence, they take a long time to generate the requested unrelated response, and in many cases they just can't help completing the sentence with a word that makes sense. As in the stroop test, but in a purely semantic domain, PFC subjects seem to be impaired in suppressing automatic responses.

Another classic test for frontal lobe functions is the *Wisconsin Card Sorting Test.* Patients are given a stack of different cards, and they must sort the cards into distinct categories (piles), according to an implicit rule (match by colour, shape, or number). Every time a card is sorted, the experimenter says if it was a correct choice or not, according to the rule. Without explanation, after ten correct responses, the tester changes the rule. A good performer will find the rules by trial and error. PFC patients fail to perform well, contrary to normals. In general, there are two kinds of errors: either they persevere in ordering the cards according to the first rule they have correctly figured out, including when the tester doesn't give positive feedback anymore, or they shift strategy suddenly, even if the tester gives positive feedbacks.

Another test, often used also in computational modelling because of its simplicity, is the *AX-CPT.* In this test, subjects observe single letters presented in sequence, and...
they are required to press one of two buttons (e.g., left) at each presentation; but they have to press the other button (e.g., right) when a target probe letter (X) follows a designated cue (A). Target pairs (A - X) are typically of low frequency, therefore the second button response is required only occasionally. The performance of the subjects depends on how able they are in maintaining contextual information. PFC patients make mistakes in this task. Either they tend to make a target response to an X probe, irrespective of the preceding letter, or they make a target response for the letter following an A, irrespective of whether the following letter is an X. They seem not to be able to consider contextual information.

There are several other kinds of tests where PFC patients seem to fail. They are organised in Table 1, according to a short interpretation of the specific cognitive capacity in which PFC patients show deficits.

<table>
<thead>
<tr>
<th>TEST NAME</th>
<th>SUBJECTS</th>
<th>DESCRIPTION</th>
<th>MAIN PROCESS</th>
<th>TESTED DEFICIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delayed response</td>
<td>Rodents,</td>
<td>The animal is trained through rewards and possibly punishments to make</td>
<td>Sensorial maintenance</td>
<td>Keeping active a short term memory of sensorial representations or intentions.</td>
</tr>
<tr>
<td></td>
<td>Monkeys</td>
<td>specific actions after the presentation of specific stimuli (usually tones for rodents and lights for monkeys) and a time delay.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delayed Object, Face, Spatial matching</td>
<td>Monkeys</td>
<td>A complex stimulus, e.g., a face, is presented, then after a delay another complex stimulus is presented; the animal is trained to indicate (through the pressing of a lever, for example) if the latter matches the first.</td>
<td>Sensorial maintenance</td>
<td>Keeping active a short term memory of sensorial representations or intentions.</td>
</tr>
<tr>
<td>Six item self ordered pointing task</td>
<td>Monkeys</td>
<td>Six different visual stimuli are presented. The animal has to indicate one of them. A subsequent presentation of the same stimuli in a different order is made, and the animal has to indicate one of them not coincident with the ones pointed in previous presentations, until exhaustion.</td>
<td>Sensorial maintenance</td>
<td>Keeping active a short term memory of sensorial representations or intentions.</td>
</tr>
<tr>
<td>World fluency</td>
<td>Humans</td>
<td>Subjects are given 5 minutes to write as many words as possible starting with a pre-specified letter (&quot;S&quot; for example), and 4 minutes to write as many 4-letter words as possible starting with another letter (&quot;C&quot; for example)</td>
<td>Verbal instruction maintenance and search in word space</td>
<td>Coming up with ideas spontaneously</td>
</tr>
<tr>
<td>Dynamic aphasia</td>
<td>Humans</td>
<td>Subjects have to invent a story from simple pictures that are showed to them.</td>
<td>Verbal instruction maintenance and search in word space</td>
<td>Coming up with ideas spontaneously</td>
</tr>
<tr>
<td>Antisaccade</td>
<td>Monkeys,</td>
<td>The subject is trained (if an animal) or instructed (if a human) to move her eyes opposite to the location in the visual field of a light point.</td>
<td>Instruction maintenance</td>
<td>Suppressing automatic response</td>
</tr>
<tr>
<td></td>
<td>Humans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequencing</td>
<td>Humans</td>
<td>Subjects are instructed to execute sequences of actions (typically arm movements)</td>
<td>Instruction maintenance</td>
<td>Suppressing automatic response</td>
</tr>
<tr>
<td>Stroop</td>
<td>Humans</td>
<td>Subjects are instructed to first read colour words, then to name the colour in which they are written (not always coincident with the meaning)</td>
<td>Verbal instruction maintenance</td>
<td>Suppressing automatic response</td>
</tr>
</tbody>
</table>
Various kinds of unitary accounts have been provided for such a variety of cognitive deficits. Shallice and Burgess 1996, for example, have proposed that the PFC is mainly a sort of supervisory system, composed of different subprocesses, involved in coping with different aspects of novel situations, in contrast to routine ones. This function requires “deliberate attention”, i.e. consciousness. “Automatic”, “routine” actions, instead, are modelled with a contention scheduling mechanism that operates by lateral inhibition of response sequences, and they can take place unconsciously. In
this model, as many others of this kind, complex unconscious executive processes are simply not possible.

Another proposal is from Duncan et al. 1996. He suggests that in PFC patients the structure of goal-directed behaviour is lost or impaired. Even when subjects can describe their goal verbally, they may stop executing their actions or may make no attempt to achieve it.

An alternative hypothesis is offered by Kimberg and Farah 1993: that the frontal lobes are used in keeping associations among the elements of WM. Damage in this area weakens these associations, lowers sensitivity to the mutual relevance of goals and stimulus attributes, and to the relations between facts and their context. Therefore, the deficits look like an executive dysfunction, but actually they are the result of damage to the available knowledge representation.

None of these unitary accounts is considered by the cognitive neuroscience community as compelling enough. Also, there is a general critique that it is possible to make of lesion and psychiatric studies of this kind. They should be taken with caution because, both lesions and mental illnesses usually involve so many areas of the PFC, or even of the entire brain, and in such a complex way, that it becomes is hard to attribute one cognitive deficit to a specific area or another. Also, in the past, scientists had to wait for the death of the patients, before they could study directly their brains. This usually occurred many years later, when the brain had deteriorated from the natural aging process.

Moreover, the lesion data, even if it is shown precisely where the damage has occurred, tell very little about what the specific brain structure does: what they say is what the brain does without the damaged structure. In fact, if a person can perform a certain task after a well identified part of his brain has been lesioned, this doesn't exclude the possibility that that part was crucial before the lesion: maybe the subject is accomplishing the task using a different strategy. And viceversa, if a lesioned brain subject isn't able to perform a task that he was able to perform before the damage occurred, this doesn't exclude the possibility that the lesioned structure was just on the path of the processing of the information between other more relevant regions. Therefore, these studies, even if generally suggestive of hypotheses and theoretical
possibilities, should be meaningfully used only as a test for theories that have been built based on other more compelling experimental findings.

3.3 Imaging Techniques

The study of humans with lesions and psychiatric disorders provides only weak evidence about the cognitive role of the PFC, but modern techniques, developed in the last twenty years, that allow us to measure where are the greatest changes of blood flow activity in the brain (PET and fMRI), are producing more convincing data to analyse.

WM tasks have been widely studied using these techniques, and the invariable result has been that the dorsolateral PFC is necessary involved for their execution. One example of this is a PET study that was conceived to confirm or falsify Baddeley's model of WM (Smith and Jonides 1997). It shows that in the brain there are different WM systems for spatial, object, and verbal information (left versus right hemisphere), that each of them has separable components for the active and passive maintenance of information (posterior versus frontal areas of the brain), and that the active maintenance is itself divided in different subareas depending on the content of the task. The results of this study seem therefore to support Baddeley's model. Other imaging studies have analysed the content of WM tasks (Petrides et a. 1993), the switching from one task to another (D'Esposito et al. 1995), and the planning of moves in a problem solving test (Owen et al. 1993). Imaging studies are also showing the involvement of the PFC in the execution of the Hayling Test (Nathaniel-James et al. 1993) and of the Wisconsin Card Sorting Test (Berman et al. 1995). In all these cases, findings have confirmed the fundamental role of the dorsolateral PFC in the active maintenance of representations in posterior areas of the brain.

The possibility given by these studies to make observations of the brain in vivo seems less open to question than lesion and psychiatric studies, but there are still substantial problems. One of them is the not fully understood relationship between the increase in blood flow and the electrochemical activity of the cells (Logothetis 2003).
Another uncertainty comes from the fact that many of the past studies involve only a very small number of subjects, whereas to have statistically significant results and reproducibility, much higher numbers would be needed.

A final point is that the spatial resolution of these techniques is still very low: these measurements can give us only, at best, a general sense of which are the brain areas involved in the execution of a certain task, but still no information can be extracted on the mechanisms involved at the cellular level. What is needed is a closer observation and the use of different techniques, at least until the imaging studies are sophisticated enough to reveal details at the single cell level.

### 3.4 Cellular Recordings

*In vivo* cellular activity of the PFC has been studied since 1960s using electroencephalographic (EEG) techniques. But what it is measured with this method is the cellular activity of huge groups of cells, and even if some results can be obtained in this way, more sophisticated techniques, that can detect the activity at the level of the single cell, have been developed only in the 1970s. These single cell studies involve monkeys and primates, studied because of the strong homology with the human brain. First, these animals are trained to perform tasks that are supposed thought to involve the PFC; then a variable number of microelectrodes are implanted in several areas of their brain, but mostly the PFC, and the electrical activity is measured while they are performing the task. The timing of the neuronal changes in activity provides insights as to which (if any) specific aspect of the performance involved the recorded neuron.

The pioneer in this methodology is J. Fuster (Fuster 1971), who also later on refined his techniques (Fuster 1989). The task in which the monkeys were involved is the delayed-response task. The animal is first shown a cue, namely the placing of a food morsel in one of two food wells; then there is a delay, and the monkey is not allowed to see the food wells; finally the animal can see the wells again, and it can also reach them. If it remembers where the food was placed, it can access it and be rewarded by eating it. Cells in the PFC of these monkeys were recorded, and the results indicated that some of these cells become active during this task only throughout the delayed
period, whereas cells in posterior areas of the brain remained active throughout the entire task, until the reward was obtained. The basic idea behind this kind of study is that some of the cells in the PFC, particularly in ventrolateral areas, are storing the information (that the food is under one of the two wells), in the form of a sustained network of active neurons. Therefore, the visual representation of the interesting well is stored in posterior sensorial areas of the brain, whereas dorsolateral PFC cells keep active this representation during the delay, and possibly also the representation of the movements to perform in order to obtain the reward. Fuster has developed a theory of the PFC, and in essence it is that this area of the brain has a fundamental role in the temporal structuring of behaviour. This is accomplished by having in the dorsolateral PFC a retrospective function of short-term memory and a prospective function of preparation for action; while the role of the Orbitofrontal PFC is to inhibit external and internal influences that interfere with the structuring of the behaviour.

Another researcher who has studied the PFC at many levels and for many years is P. Goldman-Rakic (Goldman-Rakic 1995). In her early experiments, monkeys performing delayed response tasks and control (doing nothing) were injected with radioactive tracers, which is easily trapped in metabolizing neuronal tissue, and so easy to detect with techniques that can take pictures of the radioactive tissues. Monkeys performing the tasks showed higher activity in the dorsolateral PFC, as compared with the control group.

In Fuster and Goldman-Rakic's experiments, the findings seem to suggest the following four basic kinds of neuronal patterns involved in the delayed response task: one pattern for the sensorial representation of the objects involved (wells and food); another for the motor action to perform (extend the arm and take the food); another for the reward (the anticipation of the relief of hunger); finally, the coordination of the previous three patterns, executed by keeping active the essentials of the sensorial representation, the reward expectancy and a preparation of the motor action to perform.

Further attempts to analyse which specific areas of the PFC are involved in which aspect of the WM coordinative function have been made in E. K. Miller's laboratory at M.I.T. (Rao et al. 1997). Many of his studies test the idea that different subregions
of the PFC are specialised for information regarding the form and colour identifying an object, i.e., *what* information, whereas other subregions are specialised for information regarding the location of the object, i.e. *where* information. Anatomical findings support these separate subregions, because the visual cortex has two preferential pathways to the dorsolateral and ventrolateral PFC. Their findings are that some PFC neurons do seem to be involved specifically in *what* information, others in *where* information, but it is also clear that many neurons are involved in both aspects of information (Miller 2001).

3.5 Computational models of the Prefrontal Cortex

Some points of agreement among the researchers we have seen so far are:

- The dorsolateral PFC cortex has a main role in the execution of tasks involving WM. Without it, no WM task can be executed, and when a WM task is performed, dorsolateral PFC neurons become active.

- The dorsolateral PFC seems to have a supervisory role, namely no actual sensory or motor content is present in this area, but instead it acts like a coordinator of other areas of the brain, where these contents are stored in the form of neural patterns of activity.

Some of the points of disagreement or that have not been addressed yet are:

- Does the dorsolateral PFC have internal subdivisions when dealing with different aspects of a WM task (object identification *versus* object location, visual *versus* other sensory information, verbal *versus* non verbal), or it is a more distributed network, without internal specialisations, where each neuron can subserve different kinds of tasks?
Experiments and theories on prefrontal cortex function

- Which are the specific connections with other brain areas (cortical and subcortical) essential to allow the dorsolateral PFC to perform WM tasks?

- What is the role of the neurotransmitters in the PFC?

- What are the PFC neurons computing? In other words, how is the role of WM function accomplished by the PFC? Is it inhibiting other brain areas, or coordinating them, or both?

Computational theories can address some of these questions more efficiently than general theories, because theoretical experiments can be performed, and predictions on actual experiments can be falsified or verified.

A pioneering computational study of PFC activity is the model of J. D. Cohen and his group (Braver et al. 1999). His motivation when he started these studies on the PFC was to find a model to simulate the cognitive impairments in schizophrenia (Cohen et al. 1990), which is a psychiatric condition believed to mainly involve the transmission of dopamine in the PFC. In particular he focused his analysis on the Stroop test (see Table 1), a task in which schizophrenics are known not to be able to perform well. His explanation as to why these subjects and other PFC patients seem to fail, is that there is a conflict between two processes which are 'automatic' to a different degree: one, the reading of the words, is, in a certain sense, more automatic than the other, the naming of the colour of words. The fact that the first process is "stronger" produces interference, even if it is clear to the subjects that they have to perform the "weaker" process. The role of the PFC, according to Cohen's theory, is to exert control to select which task to do: subjects impaired in this area of the brain would be unable to exert such a control, and therefore their responses will be motivated by their strength, and not by their opportunity. Cohen also built a very simple computational model of how the PFC could have this controlling role. The theoretical framework is given by the Parallel Distributed Processing approach (McClelland et al. 1988). The model of the Stroop task is a simple feed-forward network (see Fig. 8), including:
Experiments and theories on prefrontal cortex function

Fig. 8
J. Cohen's model of the Stroop Test (Cohen et al. 1990)

- an input layer with two units for the naming of the colour (red unit and a green unit),
- two units for the reading of the words (a RED unit and a GREEN unit),
- a hidden layer with four units,
- two units for the task demand (a reading unit and a colour naming unit),
- the strongest weights are from the units for the task demand, the medium weights are for the reading pair, and the weakest ones are for the naming of the colour (connections and weights are shown in the figure);
- an output layer (a red unit and a green unit).
The PFC is simulated by the task demand pool, that has the weights strong enough to balance the computations on the reading side or the naming colour side. An impairment in this part of the network has the effect of shifting the computation exclusively where the weights are stronger, in this case towards the reading process. Such a model is very simple, and it was intended just to show a minimal requirement for PFC activity.

Later on Cohen and his colleagues refined their computational framework and their model of the PFC (Cohen et al. 1996). The role of the PFC remains to adapt the behaviour of the entire cognitive system to the demands of particular task, but the model incorporates also new features. Now the simulated task is the AX-CPT (see Table 1), and the network (see Fig. 9) has a Letter Input module, with on-way connections to a Hidden module, itself connected one-way to an Output module, and also with recurrent connections to a Context module. The Context module implements the PFC activity, setting the context (AX versus not-AX), and therefore which action is required.

![Diagram of model](image)

Fig. 9
Reviewed Cohen's model of PFC activity.

This kind of model is in the family of recurrent network (Elman 1990). Other features of his model are: continuous activations (which make use of the logistic function), competitive inhibition within modules, excitation between modules, noise added to the connections between the hidden module and the Context module, and fixed recurrent weights (adjusted to fit the data from real tests on schizophrenics). An
additional important feature of the model is the simulation of dopaminergic activity: the activations in the Context units can be modified according to a gain value, which can make the logistic function smoother or sharper. Changes in dopaminergic activity induce a change in the gain of the activations of Context units: an increase in the gain has the effect of increasing the net input on the unit's activation (the logistic function is sharper), whereas a decrease in the gain decreases the net input (the logistic function is smoother), which is what would happen to schizophrenics.

Again here, as for the simulation of the Stroop test, the model is computationally very simple, and the role of the Context module, simulating the PFC, is just to balance the overall computation towards one response or another, also in the presence of noisy inputs. And again, schizophrenics, with impairments at the level of the Context module, given by the different values for the gain, would have responses based on how strong are the activations in other modules and not on the actual requirements of the task.

According to J. Cohen and his colleagues (Braver et al. 2000b), a model like this, involving a Recurrent Network, and whose basic action is cognitive control, can account for cognitive activities traditionally attributed to the PFC, namely a coordinative, a WM and an inhibitory role. In this model, the PFC is a structure that actively maintain representations of context information, and in this way it can exert a top-down control on local competitive interactions occurring somewhere else in the brain. As a consequence of this coordinative activity, the PFC can both have a WM function (to keep active and operate on representations elsewhere in the brain), and also to inhibit processes, as an indirect consequence of its modulatory effect on competitive interactions occurring in task-specific pathways.

In a recent paper, J. Cohen and his colleagues have focused on the WM role of the PFC (O’Reilly et. al. 1999). The main characteristics of the model remain the same (recurrent networks and cognitive control over other parts of the brain), but they also introduce new features. They postulate that the PFC, the rest of the neocortex and the hippocampus represent three extremes of specialisation relevant to WM: the rest of the neocortex has the role of processing information for sensory and motor activity;
the hippocampus rapidly learns arbitrary associations; and finally, the PFC keeps active (or inhibits) and operates on the other two systems.

Especially interesting is the introduction of a role for the hippocampus in the WM processes: its role, as described also in previous connectionist models (O'Reilly and McClelland 1994), is to perform pattern separation, when inputs are considerably different, and pattern completion, when the inputs are considerably similar. They also suggest that the PFC is not organised according to intuitive dimensions (like the what and where distinctions of E. Miller), but instead it has more abstract modalities. WM is interpreted as the active portion of the long-term memory, which itself is distributed over many brain regions.

A different model of PFC activity can be found in Dehaene and Changeux 1997. They view planning as the main higher function that the PFC is performing. Their hierarchical model, tested on the Tower of London (see Table 1), consists of a “gesture” level, implementing the sensorimotor coordination needed to point to the location of the pegs, an “operation” level, where each operation is a sequence of two elementary gestures, and finally a “plan” level, which implements a trial and error exploration of alternative moves to go from a start state to a goal state. The last level implement the functions of the PFC. It has WM units, that maintain memory of a previous problem when a plan is tested; plan units, that generate new plans by activating novel patterns among lower level operation units; and reward units that evaluate the success or failure of the plan.

This model, although capable of matching some of the real data on normal or impaired subjects performing the test, has several important limitations such as no learning ability, and no action chunking mechanisms. The importance of their model is in its attempt to explore the planning interpretation, and an illustration of the neural circuits that can code for subprocesses of a general planning role, namely the working memory, plan generation and rewards.
3.6 Towards a unified view of PFC function

The striking number of different descriptions that can be found in the literature about the role of the PFC gives an indication of how elusive this region of the brain is: storage, execution, attention, inhibition, problem solving, decision making, multitasking, just to mention a few.

In the very recent "An Integrative Theory of Prefrontal Cortex Function" by E. K. Miller and J. D. Cohen 2001, there is an attempt to unify these descriptions. I review this article in detail here because the general lines they develop are fully in accord with the model I develop in the next pages.

The authors argue that the main role of the PFC is to represent goals and rules of the behaviour, via an active maintenance of patterns of activity, and to configure processing in other parts of the brain in accordance with current task demands (top-down signals), especially when less habitual responses and high flexibility are needed. Repeated selection of responses can strengthen specific pathways, that can become independent of the PFC. If an unfamiliar situation is presented, a search starts for the best behaviour to achieve a goal, until it is met with success: then reinforcement signals augment the corresponding pattern of activity by strengthening connections between the PFC neurons activated by that behaviour and those representing the novel situation. Minimal requirements for a theory of PFC function are:

- Robustness: the PFC must maintain its activity robustly against distractions until a goal is achieved.

- Flexibility: it must be flexible enough to rapidly update its representations when needed, according to task demand.

- Representational adequacy: it must house the appropriate representations that select the neural pathways needed for the task.

- Integration: it must integrate stimuli, internal states and responses.
- Plasticity: it must be highly plastic in order to acquire new goals and means.

The main properties of the PFC are:

- The PFC is the locus of convergence of sensory inputs from the rest of the cortex, limbic connections associated with internal states, motor outputs, and intrinsic connections.

- The PFC neurons become, through learning, selectively responsive to specific cue-reward associations or more complex rules (like if-then).

- The PFC is capable of sustaining activity in the face of intervening distractions.

- Feedback pathways from the PFC can exert biasing on other brain structures.

- The PFC is capable of learning rules involving associations between stimuli that are separated in time: this might be achieved via Dopaminergic neurons that seem to be coding "prediction" errors.

In simple computational models, the PFC units influence the activity of other units whose role is carrying out the required input-output mappings. Therefore the PFC decides, by keeping active certain patterns, which pathway in the rest of the cortex must win and perform a certain behaviour. Storage, execution, attention and inhibition are all unified in the view of the PFC as the locus of representation of the selections of the desired combination of stimulus features to be mapped onto the response over other competing combinations. The hippocampus has the role of binding together information into a memory of a specific episode, whereas the PFC extracts the regularities (rules and goals) across episodes. The Hippocampus implements “weight-based” control: it helps to consolidate permanent associative links between the pieces
of information that define a long term memory. The PFC is instead activity based, highly flexible, and its ongoing activity specifies the pattern of neural pathways that are currently needed.

Two demands must be satisfied by representations in the PFC at the same time: they must be responsive to relevant change in the environment (flexible), and they must be resistant to irrelevant changes (robust). This double demands might be conflicting, and they need a regulatory mechanism. Dopamine may play a role in gating, by modulating the influence of its afferent connections, and essentially establishing when representations must be robustly maintained (gate closed), or when they can be rapidly updated (gate opened).

Questions that, according to the authors remain to be experimentally and theoretically analysed are:

- Can this system support subgoaling or hierarchical planning?

- What is the representational power of the PFC: in particular, how can its compositional power and subgoaling capacity be modelled by neural networks?

- How is the PFC functionally organized: e.g., are there what versus where cells in this area of the cortex?

- It seems that anterior cingulate cortex allocates how much control should be produced by PFC, by modifying the intensity of its pattern of activity, and in this way it detects possible conflict in processing. But the causal relationships in this process are still unclear.

- The maintenance of activity in PFC neurons has been modelled in two ways: cellular (triggered by inputs and maintained via the activation of specific voltage-gated channels) or circuit based (through recirculation of activity in recurrent loops, or as attractor networks). But it is still not clear why active
maintenance is so severely limited in its capacity. For example, it seems to work serially. No theory has provided a model where this limitation emerge.

- How can long term plans (one day or more) be explained? It seems that active maintenance is not plausible: more likely they are stored somewhere and then activated at the appropriate time. Maybe the hippocampus, so fast in learning, might also be involved in this process. This structure might encode an association between the desired goal representation within the PFC and features of the circumstances under which the goal should be evoked. The detailed nature of such interactions, their relationship to the dopaminergic gating and learning mechanisms just described, are still to be determined.

In this thesis my intention is to address in detail the first three of these questions, and to set all of these issues in a coherent framework concerned with planning.
Chapter 4

A framework for neural planning

In this chapter I describe an abstract model of neural planning (see also Fig. 10, at the end), as taking place in the mammalian brain, consisting of a circuit involving mainly the PFC, the basal ganglia, posterior cortex and the cerebellum. Its computational modules are motivated by Reinforcement Learning requirements for planning, that will be extensively illustrated in chapter 5. Each of these modules has its biological counterpart, i.e. an area of the brain that experimental evidence indicates may have a function analogous to the role of the computational module. I discuss these neural counterparts in the form of hypotheses, that will be more extensively illustrated, motivated, and tested in subsequent chapters.

Note that an implementation of this abstract model can be a connectionist network, a Gandy Machine (De Pisapia 1999), or any other computational device that follows the general design. The goal here is to show core computational requirements for reward based planning modules, and plausible neurobiological counterparts. The actual computational device would matter if the main concern were detailed biophysical plausibility, but this is not the main focus in this thesis. Here I give a description of the role of the PFC at the computational level, according to the division into three levels (computational, representational and implementational) in the study of brain and cognition, due to D. Marr (1982), and summarized in the Introduction.

In the next chapters I also provide descriptions at the representational level of how the computational task of reward based planning can be achieved, e.g., using Synchronous Value Iteration and Temporal Difference learning (Sutton and Barto
A framework for neural planning

1998). What algorithms might be closer to the actual neural implementation is still an open question.

Neural planning is a complex activity, and I do not exhaust the topic. The computational description of how the PFC, in concert with other brain areas, achieves a planning capacity comprises a number of hypotheses. I first illustrate them briefly in this chapter, and then I discuss them in the next chapters, showing the computational justification and the plausible neural counterpart. I will finally make some comments on their testability in the last summarizing chapter.

=================================================================================

The main thesis asserted in this dissertation is that unconscious planning is implemented in mammals in specific brain areas following Reinforcement Learning principles and Temporal Abstraction. This thesis is composed of the following main hypotheses:

4.1 Hypothesis 0 – Unconscious and conscious planning

There is a basic adaptive capacity of the mammalian brain, which I will call indifferently unconscious, implicit or neural planning, to simulate and evaluate goal oriented behaviour and consequent experience, in order to possibly execute the best (given current knowledge) sequence of actions, that should be performed in order to achieve a goal. This process is not the same as free will, conscious decision making, inner goal oriented speech (Vygotsky 1986), or the following of social instructions on how to achieve goals, even though these complex processes are possibly based on it. Consciousness is not required for implicit planning, although the former may have access to part of the latter. Implicit planning is a form of information processing, implemented in the single brain, and consciousness of (part of) it is a much more complex mechanism, that takes place only when social factors, communication with other members of the same species, or an internal simulation of this communication process (the inner speech) are involved.
I discuss this complex topic, with several philosophical implications, in the next chapter.

4.2 Hypothesis 1 – Planning and action chunks

Implicit planning is achieved through processing of information that is based on reinforcements, and on the chunking of those sequences of actions that are several times successful in reaching a goal/reward. This chunking process facilitates and speeds up the planning process, because chunks normally don’t need to be re-evaluated, but they are well acquired knowledge about how to go from one state of the environment (start) to another (end/goal/reward).

Formation of chunks for actions, that allows faster and faster planning, takes place in the basal ganglia, particularly it has been studied in the dorsolateral striatum. A preliminary selection of available options, i.e. chunks of actions, takes place at the level of the basal ganglia through a competition mechanism. Resolution of conflicts between concurrently valid options is performed by anterior cingulate cortex. Orbitofrontal cortex is instead engaged in the evaluation of the several remaining options.

I explore and test this hypothesis in chapter 6.

4.3 Hypothesis 2 – Planning and internal model formation

Implicit planning is achieved through phasic dopaminergic effects from ventral tegmental area, acting in the PFC as an error signal in the prediction of every task related stimulus, thus allowing the construction of the internal models. Novel internal models of task relevant information on the environment are thus constructed and maintained in the dorsolateral PFC in the form of activation based processes. In these internal models information about where, what and when goal related events is learned and put together.

I discuss at length this hypothesis and test with simulations in chapter 7.
4.4 Hypothesis 3 – Planning and internal model storage

Lastly, posterior cortex (i.e., neocortical areas other than the PFC) and cerebellum consolidate internal model representations of the world once the activation based models in the PFC have been several times successful in their predictions. This is a key computational passage that allows efficient planning. It is the store of knowledge of execution of well acquired (sub)tasks, so that, when evaluating new courses of actions, the steps and the state transitions stored do not need to be recomputed, but only information about the start state and the end state is used. In the brain, I have said that it is the PFC’s task to construct new internal models, and it is the basal ganglia’s task to label chunks of actions related to forming these internal models. Therefore, it remains to discuss where the successful and iteratively used internal models are stored, once they don’t need to be learned anymore. My proposal, hypothesis 3, is that they are stored in the posterior part of the cortex, the cerebellum and the posterior part of the basal ganglia (which come into play for this storage role, that is to say when no learning is needed). The idea that the anterior part of the cortex is need more during learning, as opposed to the posterior part which is used more when the execution of a task is more automatic, is not new. However, this idea is often not used when studying local cortical areas. Overall cortical and subcortical dynamics should be taken much more into account when focusing on local areas. Neuroimaging studies showing how different brain areas are engaged while performing the same kind of task depending on if subjects are in a learning or in an automatic phase (e.g., Jueptner et al. 1997a, Shadmehr and Holcomb 1997) will be discussed in the last part of chapter 7. Even though this hypothesis refers to a key and conclusive step of the planning process, I will not discuss it at length because it doesn’t involve primarily PFC activity.

4.5 Computational framework for implicit planning – Modules

The overall computational framework for implicit planning is shown in Fig. 10, together with the corresponding neurocomputational hypotheses. An internal model
module (dorsolateral PFC) updates multi-time internal models of the world and maintains them in an active state. The job of the units in this module is to code information about what, where, and when the goal of an higher level action will be present in the environment.

An anticipated rewards module (orbitofrontal and anterior cingulate cortex) actively maintains goal representations and therefore evaluates task relevance of different chunks. A chunk labeling module (striatum in the basal ganglia) incrementally chunks the actions, under the directives of the previous modules, and single cells code for the pair start/goal, defining the chunk. The inputs are given by high level information about space and object identity in the perception states module (parietal and temporal cortex). A prediction error module allows learning by sending an error in the prediction of reward occurrence (to the striatum) and of task related stimulus occurrence (to the dorsolateral PFC). A reward signal is sent at the time of occurrence of a reward to the chunking module. Finally an actions module stores the actual
actions that define each chunk (motor and premotor cortex, but also the cerebellum, not shown).
The connections in this framework are biologically motivated, and what they actually are in simulations depends on the specific way the modules are implemented. The actual implementations I do in this work, which test only subcomponents of the general model, are simple connectionist networks as in Montague et al. 1997. The units are binary, and they can represent single stimuli, single actions and rewards. Time is represented as in Complete Serial Compound, i.e. stimuli, rewards and actions are represented as different units for each different time step. Much of the information is stored in the weights, whose values are determined by learning processes based on Temporal Difference learning and Temporal Abstraction, in the form of the equations given in section 3 of chapter 6. Specific units compute the difference in the expectation of the reward delivery at two consecutive time step, according to the Temporal Difference definitions.
Chapter 5

Hypothesis 0: unconscious planning

The real continuous process takes place below our consciousness: the series and sequences of feelings, thoughts, and so on, are symptoms of this underlying process.

Consciousness in general developed itself only under the need to communicate

Friederich Nietzsche, *The Gay Science* (1887)

In this chapter I discuss the relationship between human planning, language and consciousness, and if there is such a thing as implicit planning. This topic is extremely complex, but nonetheless here I want to indicate the main questions involved and the need for further experiments to find answers. My aim is to point the attention of the reader to the fact that human planning in completely new situations can also be a non-linguistic and unconscious process, in contrast to the hypotheses underlying most of the experiments in cognitive neuroscience. The biological neural mechanisms that allow the generation, evaluation and selection of actions can follow ways that are completely unaware to the subjects. My claim is that the study of unconscious planning will bring a wide new set of insights into higher cognitive functions. For this, a new experimental framework appears necessary.

5.1 Motivation

The topic of unconscious planning is a strange and difficult topic. At first it may appear simple, because many people immediately would agree that the way in which we organize our behaviour, motor or cognitive, is not completely known to ourselves.
After all, it is a common experience that we do quite complex things without actually being able to describe all the steps that we do, or where we take them from, as when lacing our shoes, when structuring phrases in our natural language, or when driving home on the usual road. On the other end, when we consider our behaviour, or the behaviour of other people, we tend inevitably to attribute intention, consciousness and will. The idea that our choices and decisions might be following to a significant extent fully mechanical and unconscious processes is deeply challenging to the way we interpret ourselves and others. We do not see ourselves and others as if we were puppets in the hands of physical and social laws. This would be an unacceptable paradox, simply because we cannot be a stranger to ourselves. The perennial problem of freedom of the will presents itself in full.

Any research touching the question of consciousness faces “hard” problems (Chalmers 1997), such as the problem of how to measure consciousness, of how to treat qualia, and if it makes sense to discuss subjectivity in scientific terms or not. I won’t discuss directly these profound problems here, except for a mention of how to measure consciousness. I see this core problem related to the fact that on one side consciousness is a subjective phenomenon, that therefore cannot be scientifically studied, but on the other hand it seems that consciousness has an influence on objective events. I don’t think any scientist has denied the existence of the subjective aspect of consciousness, even the most strict behaviourist, but many have been in disagreement on what to consider its objective counterpart, the part that can be studied. In other words, what is the causal role of consciousness? Related to this, is the problem of the measurability of consciousness. In order to conduct scientific research on consciousness, some way to measure it, or to indicate when it is present or not, is needed.

My approach is given by the following definition of how to measure consciousness: if an agent reports verbally, e.g., through protocol analysis, about some information, then we can say that the agent is conscious of that information. Verbal reports are the standard operational indexes of consciousness (Baars 1988). To be more precise, this is actually a semi operational definition, as discussed in Dennett 1969, and it consists in identifying in what agents say as one of the objective counterparts of
Hypothesis 0: unconscious planning

consciousness. This definition is approximate, and it gives only a sufficient condition for consciousness, but it doesn’t say anything about necessary conditions. This definition therefore doesn’t exclude that non speaking entities, like animals, might be conscious, but there is a problem of accessibility to such consciousness, whose subjective components are precluded to an observer. It might be claimed that this semi definition also does not give a sufficient condition, because machines and robots might be capable of reporting their internal states, and this would not suffice to confer consciousness. A full discussion about this objection would lead us too far from the main topic here.

Despite all these difficulties, some discussion of consciousness is necessary if you consider the tests of human planning in the cognitive neuroscience literature, from the perspective of those who are trying to construct computational models of human PFC activity. Most, if not all, of these tests (as summarized in Table 1 in chapter 3) view planning as a quite elaborated process, that uses language, full of concepts and strategies, or, more importantly, with declarative and explicit instructions given from the experimenter. Mostly, subjects in these tests are supposed to be conscious of what they are thinking and planning about. Usually they are asked to write verbal reports on which are the strategies that they have used to attempt to solve their task, and this is considered all that there is to know about the decision process. In these experiment, subjects are given explanations of what to achieve and a basic degree of how to achieve it (rules of the test).

But why do we consider planning and goal oriented behaviour in humans as a primarily conscious activity? There is undoubtedly a conscious side of these activities, indeed the one we report to ourselves and others, but why assume that this is the fundamental aspect of planning? In computer simulations there is no meaningful way to make use of these tests, because the focus must be the planning process in its most basic and simple form, i.e. a basic neural process that can achieve and improve hierarchically structured behaviour once the current state of the environment and the goal state are given as inputs. Conscious planning seems to be, in neural terms, a much more complex phenomenon.
Consider a commonly used cognitive neuroscience manual, Gazzaniga et al. 1998. In chapter 11, on executive functions and frontal lobes, there is an identification of complex organization of behaviour and intact PFC function with the capacity to choose and to want, both considered as conscious activities. For example, we find in the same paragraph these sentences (page 453-4): “We choose to act because we want to accomplish goals, to gratify personal needs”; “the ability to form a coherent plan of action is compromised after damage to the prefrontal lobe”. The sense of the paragraph is that goal accomplishment is equivalent to willed behaviour, and the prefrontal cortex is its neuroanatomical counterpart.

Another manual, Zigmond et al. 2002, is even more explicit (page 1397) “All normal people have a strong subjective feeling of intentional or voluntary control of their behaviour. Asking people about goals or intentions is probably the single most effective way to obtain information that is predictive of their behaviour during problem solving”.

Take the model of attention for action developed by Norman and Shallice (1986). It is assumed that the central executive function, for which the PFC is responsible, is required when a response is complex. This function requires “deliberate attention”, i.e. consciousness. “Automatic”, “routine” actions, instead, as we have seen in chapter 3, are modelled with a contention scheduling mechanism that operates by lateral inhibition of response sequences, and they can take place unconsciously. In this model of attention, as many others of this kind, complex unconscious executive processes are simply not possible.

Consciousness has been considered for decades as a sort of taboo or distant topic for science, and therefore scientists traditionally have translated subjective accounts of cognitive activities into third person accounts (as in information processing studies of vision, or motion). But when it comes to high level planning first person accounts and consciousness become instead the norm, and the cognitive implicit planning capacity is not considered at all.

It might be objected that I am looking for unconscious planning because I haven’t a model of consciousness, and so I wouldn’t know how to model conscious planning. And this is true. But the point is, if experimental evidence shows that, given the above
Hypothesis 0: unconscious planning

operational definition of consciousness, we can find that people organize their behavior, create and evaluate complex strategies without being conscious that they are doing it, then this is a phenomenon that deserves attention, at least as much as the conscious strategies reported in verbal protocols. The idea here is therefore to see if an unconscious, perhaps more basic, planning capacity can be found in the whole human brain. It would not necessarily be declarative and linguistic (although it can be, and still be unconscious), and people would not be able to make a verbal report on it. Such planning would be a non instinctual process, that is not a hardwired fixed behavioural pattern, but instead a mechanism open to learning of new strategies, of new ways to evaluate and correct options on the basis of previous experiences and goal achievement, without necessarily mirroring the behaviour of other members of the same species, and without necessarily adopting linguistic instructions on how to do things.

The idea that automatic movement in humans can be performed without conscious control is somewhat familiar to healthy subjects. When we drive on highways, to cite a classical example, for very long distances we might be completely unaware of what are the actions we have performed, how was the landscape, how many times we have overtaken, and so on. Nonetheless several complex actions have been taken, and several decisions and evaluations also, but they were within a range of well known situations, and so we could keep on focussing on other things. Similarly we accept the idea that skilled players in complex games, like chess, might spend most of their time on high level situations, and they can leave the solution of simpler cases to an automatic process. Several years of studies on this mechanism were spent by H. Simon and A. Newell at the end of last century. Think for example when you have learned the rules of a new game: it’s on that rules that you spend most of your conscious effort. But when these rules are acquired, you don’t think about them anymore, because you concentrate on simple strategies, and the components of the game start to be perceived differently. For example, in chess, as you look at the king you have to see it as the “slow mover”, and you don’t need to repeat consciously to yourself the rule that constraints its movements.
Therefore we easily accept the idea of an unconscious organization of behaviour for motor and cognitive skills, but what I am searching here is something more. My question is if we can find an unconscious organizer, learner and planner also in completely new situations. The traditional view in neuropsychology is that the learning of a skilful motor behaviour is initially conscious, possibly under verbal control, but then, as a function of training, this conscious process opens the way to an automatic processing. Implicit memory of motor skills is therefore well acknowledged, but it is believed that it only takes place when learning is finished (Popper 1972, or Halsband et al. 1993). A similar position is held in the language domain. Although the actual way in which the sentence is formed, and the articulatory movements we perform with our mouth to produce meaningful sounds can easily be seen as unconscious and automatic, the high level process of planning and choosing what to say remains intentional, willed, and conscious (see Schaffer 1984 for example).

But what if there is also a simultaneous unconscious process that is learning and planning cognitive and motor activity at a high level? Planning in new and complex situations is traditionally thought to take place only with conscious effort, and to the automatic process it is left the well known way to execute skills. But what if people tell in their verbal report on how they solve problems or choose to act is only part of the full story, in the sense that new and crucial elements have been learned and evaluated unconsciously? Consciousness might be covering an underlying unconscious motor and cognitive learning process. Or it might be a retrospective explanation of unconscious processes.

Language and consciousness involve several social aspects that would make sense to model in connectionist terms if take into account several planning agents together, and how they can collectively achieve goals, by means not of individual neural communication (the spikes), but by some kind of natural language. As the etymology of the word “conscious” suggests (the Latin *cum scire*, meaning “know with”), and also our semi operational definition, the ancient Romans denoted as “conscious of something” someone who could share knowledge of this something. Consciousness is a social phenomenon, that is to say a phenomenon that involves several brains.
Hypothesis 0: unconscious planning

together, and their capacity to communicate. I would claim that consciousness in a single brain always involves the individual capacity to model other minds, and so to simulate an internal discussion with them. Similar ideas on the social role of consciousness can be found in Barlow (1987), in studies of how internal speech develops in Vygotsky (1986).

Language can of course work very efficiently as a planning mechanisms, as when we reason with ourselves about how to reach a certain target (of any nature), and make use not only of personal experience, but also of instructions given to us within our cultural framework, which is indeed the framework that shapes most of our actions. But the focus of the cognitive research should be on a more basic form of planning, that doesn’t necessarily use language, symbols, explicit rules or social inputs. The number of concepts that are shared with other individuals when solving tests like the Tower of Hanoi, for example, is enormous: things learned at school, on books, or talking with others. An unconscious planning process might be behind these linguistic capacities, and it might be shaping the social planning capacities of the single agent.

Cognitive neuroscience has been studying planning mostly as a social phenomenon, but looking at if it were an individual one, whereas it should study the biological capacity of a single brain to code information in such a way to simulate, improve and select high level actions.

Would such an unconscious process involve the PFC, or just subcortical structures? Most of the researchers who talk about planning do seem to have the view that they are the basal ganglia that are responsible for automatic and implicit organization of behaviour, whereas when the PFC is active, then necessarily the individual is conscious of how and what she is trying to achieve.

The PFC cortex, together with the temporal lobes, is certainly involved in language processes, and in consciousness, but my fundamental question here is if the PFC is primarily involved in non linguistic implicit planning and decision making. My question is if humans can plan, i.e. generate and evaluate strategies to achieve goals, without being conscious of these strategies and goals, indeed without the capacity to make a verbal report of either the options they are generating and evaluating, or the goals that they are trying to achieve, and without necessarily the use of social biases.
Hypothesis 0: unconscious planning

Are there experiments in cognitive neuroscience on purposeful unconscious human behaviour? The answer is: not many. Studies on animals go in the direction of considering behaviour without consciousness, because obviously we do not expect from them verbal reports. But the two possible modes of planning, conscious and unconscious, are in a certain way even more confused in these studies, because we have no methodology to claim in which mode the animals are. For example, as we have seen, there are plenty of studies on working memory in monkeys, but how they deal with the problem of how this kind of memory relates to consciousness (one of the themes discussed in Miyake and Priti 1999) is highly unclear, because, surprisingly, this issue has not been considered much.

Apart from planning, the idea of unconscious activities has instead become quite acknowledged in cognitive neuroscience, and implicit cognition has gained a status of its own in recent literature. For example, subliminal perception is a well known phenomenon, and it occurs whenever stimuli presented below the threshold of awareness are found to influence thoughts, feelings, or actions. The term subliminal perception was originally used two centuries ago to describe situations in which weak stimuli (like very low sounds or distant shapes) were perceived without awareness. More recently, the term has been used to describe any situation in which it can be showed that unnoticed stimuli are nonetheless perceived. Examples of subliminal perception are found in patients with neurological damage, who claim that they do not to see particular stimuli, but nevertheless they respond on the basis of information conveyed by these stimuli.

One example of such a syndrome is blindsight. Patients with blindsight have damage to the primary visual cortex. As a result of this damage, they are often unaware of perceiving stimuli within a restricted area of their visual field. But even though they may claim not to see stimuli located within the "blind" area, they are still able to guess the size, shape or orientation of the stimuli that they claim not to see.

Another example is unilateral neglect, which is associated with lesions of the parietal lobe, typically the right one. The patients behave as if the left half of their visual world were missing (if shown a picture of a house and told to copy it, they draw only
Hypothesis 0: unconscious planning

its right half). Even though the subjects neglect the left half of their visual field, it can be shown in some cases that the brain is processing missing information. Another syndrome in which subliminal perception occurs is prosopagnosia or face agnosia, which typically involves lesions in the temporal lobes. Patients with prosopagnosia are unable to recognize familiar faces. Although they may be aware that they are looking at a person's face, they are unable to say who the person may be. Thus, prosopagonsics have no awareness of perceiving any information regarding whose face they may be viewing. However, despite this absence of awareness, some patients with prosopagnosia are able to choose which of two names goes with each familiar face that they claim not to be able to recognize.

Perception without awareness of perceiving can also occur in surgical patients undergoing general anaesthesia. When memory is assessed about specific stimuli presented during the anaesthesia, there appears to be some memory for such presentations.

These are well acknowledged examples of unconscious perception, but what about unconscious actions: is there a brain disorder that involves the inability to perform actions consciously, but that nonetheless leaves the capacity to perform the same actions unconsciously? This disorder indeed exists, and it is called apraxia (Brown 1988). It is usually unilateral, and occurs mostly in patients who have suffered some kind of brain damage to one hemisphere, due to stroke or accident. If these patients are requested to perform a movement with the affected side of the body (e.g., the right hand), like combing their hair, they are unable to do so, whereas they can with the unaffected side (e.g., the left hand). Besides, they can also perform the movement if they are asked to copy another person doing the same movement, or if they are left to perform the movement in an automatic mode, namely for an “internal need” or in usual situations. This disorder is still far from being fully understood, and the relevant literature is confused, but some attempts to construct a theory of it have been made since the beginning of the last century, in particular by Liepmann. He classified the illness into three groups, on the basis of the different symptoms: limb-kinetic apraxia, involving incompleteness of even the simplest motions, exchanges of motion, and imitation and spontaneous performance are both affected; ideomotor apraxia, where
Hypothesis 0: unconscious planning

under some circumstances simple motions can be performed, but not when the patients want to or is requested to; in ideational apraxia, finally, where the simplest motions are performed well, like lighting a match, but sequences of simple motions are impaired, in the sense that there are omissions, wrong ordering, actions are performed before than they should, or the right actions are directed towards the wrong object.

With regard to the specific areas of the brain affected, since the early studies from Liepmann, it has been found that for right handed subjects the left hemisphere seems to be the repository of the “movement formulae” that control purposeful skilled actions of both sides. The role of the right hemisphere is less understood, and although it does exist, seems to be less context dependent than the left. Apraxic patients that have a problem in conscious control, could be easily explained as having intact the storage place for automatic actions, so if they are triggered not consciously, they can be performed.

What this disorder tells us within our discussion is that indeed there appear to be two ways in which actions can be performed, namely an unconscious and a conscious one. Whether these are two extremes of a continuum, and if the brain areas involved are separated or not, remains to be understood, but the general idea that there are two modi operandi for actions seems to hold.

In conclusion of this section, I claim that because cognitive neuroscience experiments have been biased towards conscious aspects of planning, implicit aspects have been largely ignored. We have seen apraxia. It is a disorder that involves the impairment of unfortunate patients to execute willed actions. And therefore, I argue, given the assumption, experimenters have only checked the planning capacities of these people in novel situations by giving them instructions, which necessarily involve consciousness. But why not test them by giving complex tasks without telling them what the task is about, but by rewarding them if they reach the goal? If an unconscious planning process exists, if it can be somehow physically dissociated from the conscious one, those lesioned patients who have the unconscious system preserved could perform well in these tasks, improve their performance, and not consciously know anything about it.
By definition, the way in which we access the activity of our own brain is through consciousness, whatever that is. We accept that part of our brain activity is non conscious, but we cannot have conscious direct knowledge of the unconscious, because it would be a contradiction. Therefore we simply assume it on the basis of observed behaviour, ours and those of others. But this assumption is quite artificial and hard to incorporate in our interpretations and in the way we test our interpretations, including scientific experiments.

In the next section I review data that indicates that implicit planning is a mechanism that exists or at least deserves more testing.

5.2 Experiments on unconscious learning and decision-making

Experiments in cognitive neuroscience on unconscious planning are rare, but the idea that complex behaviour might be unconsciously organized and learned is not new in philosophy. For example see the book “The discovery of the unconscious before Freud” by L. L. Whyte (1960) as a starting point, or more recently “Consciousness and the computational mind” by R. Jackendoff (1990).

In psychology, in the first part of the last century, the idea that people solve problems and make decisions for reasons different than those that are conscious to them seems to have been known only to Freud and the followers of the psychoanalytic tradition, like Anna Freud or Edward Bernays. The literature is sparse until the 1970s (Baars 1986). It was only then that a few psychologists started to experiment on people viewed as irrational decision makers, and as solving problems on the basis of information that they were not aware to have. Kahneman et al. (1983), Nisbett and Wilson (1977) or Langer (1975), for example, were separately showing how people sincerely justify their acts according to causes quite distant from the real ones, similarly to what happens in confabulation. For example, Kahneman showed how even experts in the field, when presented with three different socks in random sequence, if asked to judge the best one for quality, indicated the third in the series well above chance. The preferred sock was different for each expert, and they justified their choice with quality considerations, nonetheless what really remained
constant was an inclination towards the last presented sock. Kahneman, Slovic and Tversky 1983 continued these studies showing that rationality or logic are largely independent of decision making, and people (including statisticians) often substitute non-optimal choices. Similar research takes place today, especially in marketing studies, but also in pure psychology. In this respect, it is especially important the work of J. A. Bargh, as reported for example in The Automaticity of Everyday Life (1997), where he argues that most social behaviour is automatic, and it occurs in response to an environmental trigger that works as an unconscious prime. He claims that almost all mental representations are constructed automatically and unconsciously (see also Bargh and Chartrand 1999). In Berry and Broadbent (1984) subjects interacted with a “computer person”, which gave responses following simple or complex rules. Simple rules could be learned and made explicit; complex rules could be learned implicitly. Another set of experiments concerning the phenomenon called implicit learning is more relevant to planning, and therefore I discuss this more in detail. These studies, started by A. S. Reber (1993) in the 70s, seem to support the idea of an unconscious capacity to learn structures in the environment and to shape behaviour according to them.

One way to assess implicit learning of sequences of stimuli and actions is to use artificial grammar tasks. In one version of these experiments, the serial reaction time task, participants are shown apparently random sequences of letters (e.g., TSSXXXY). After the occurrence of a letter, the participants are asked to press a corresponding button, one for each letter. Unbeknownst to the subjects, the letters are actually showed following sequencing rules, as in artificial Markovian grammars (see Figure 11). For each grammar the transition probability from letter to letter can be a value from 0 to 1 (in figure 11 I only show the letter that is generated when going from one state to another, without the transition probabilities).
It has been repeatedly shown that, while doing this task, subjects improve their performance (i.e. their reaction times get shorter and shorter) if the grammar is kept fixed, but, as soon as the underlying grammar is changed, again unbeknownst to them, their reaction times become worse (i.e., longer) than at the beginning of the task. This is considered a proof that humans can improve their performance by learning about structure (grammatical, in this case) without becoming conscious of what they have learned, or that the structure has changed. Another way to assess implicit learning is to ask subjects to memorize a set of letter strings generated by a Markovian grammar, and then they are asked to say if they think that a further set of novel strings follows the same grammar or not. Subjects generally perform above chance on this task, but they are unable to report the rules of the grammar involved and to explain how they decide.

Studies on implicit learning are still a very active field of research (see Reber 1993, Destrebecqz and Cleeremans 2001). They also attract several criticisms, as for example in Shanks and St. Johns 1984, who introduced some criteria by which to judge methods in similar experiments. I won’t discuss these criteria here, but just cite the fact that these studies assess consciousness using post experimental verbal reports,
Hypothesis 0: unconscious planning

whereas a more careful method that would satisfy these criteria would be to make subjects report during the training. One of the conclusions of Shanks and St. John is that simple information is enough for these subjects to succeed above chance in their learning tasks, even though the rules might be complicated; moreover, this simple information is reportable verbally and thus is conscious. Therefore these authors claim that in these experiments no form of learning has been proved to occur without consciousness, but just with awareness of pieces of the whole set of rules involved. They make the extreme claim that no learning would occur without consciousness.

One of the interesting answers to this criticism (Destrbecqz and Cleeremans 2001) have pointed to the consideration that conscious and unconscious modes are not just the two options available, but they can represent the extremes of a continuum. The underlying learning mechanism is the same, but if not enough information is available, or if not enough time is given to subjects to form explicit representations, then more implicit representations are formed. The authors tested these claims with a new set of experiments, called Process Dissociation Procedure, where subjects are given different times to learn Markovian grammars: by reducing the time available for processing the stimuli, the experimenters impaired the development of explicit representations, but left the capacity to form weaker memory traces.

Another way to modulate the level of consciousness of the hidden structure would be to make the transition probabilities from one letter to another as close as possible to 0 or 1 values (e.g., an X is always followed by an L and never by an R). This makes the grammar close to being deterministic, increasing the likelihood that the grammatical rules will be appreciated explicitly.

The debate on these issues is very much active, and I do not expect to be exhaustive here, but if we look at the brain regions that are involved in these putative mechanisms, we find considerable support for our general view of unconscious planning taking place in the PFC. A PET (Positron Emission Tomography) study has been performed (Berns 1997) on the previously described serial reaction time task. Some rudimentary planning is involved here, in the form of learning sequential dependencies among stimuli, and using these to anticipate the next stimulus in order to improve performance. The results indicate a major involvement in this task of the
ventral striatum in the basal ganglia with regard to novel information (grammar change), and right dorsolateral PFC with regard to maintenance of contextual information (grammar transition probabilities). This study was limited to subjects who were never aware of the fact that the occurrence of the letters was driven by grammars, and therefore full consciousness never developed.

Another way to assess if subjects can improve their performances without knowledge of it can be found in the study of Bechara, Damasio et al. 1997. They have investigated the hypothesis, very much in line with what I have tried to expose here, that fully conscious declarative reasoning might be preceded by unconscious biasing “steps”. They also further hypothesized that the unconscious activity might be based in neural systems other than those that support conscious and declarative activity. They tested their hypotheses on normal and prefrontally impaired patients an a gambling task (the Bechara task), consisting of borrowing 2000 facsimile U.S. dollars, and choosing cards from one of four decks (A, B, C and D). The choice of a card resulted in a gain or a loss, according to untold rules. The rules consisted of a loss of 250$, 1250$, 50$ and 250$ respectively for the four decks, with a respective probability of 50%, 10%, 50%, 10%. In the remaining cases, there was a gain respectively of 100$, 100$, 50$ and 50$.

Once you know the rules, you can see that the advantageous strategy in the long term is to choose as much as possible from decks C and D, but the players in the actual test had no explicit knowledge of the rules, and given the probabilistic nature of the task, it might take a lot of time before the explicit knowledge is formed. Nonetheless the experimenter found that the normal subjects began to choose advantageously before knowing which was the best strategy (whose ignorance was assessed with verbal reports), whereas the prefrontal patients kept on choosing from the disadvantageous decks, even after being told explicitly which was the advantageous strategy. The unimpaired subjects also developed anticipatory skin conductance responses when they where evaluating a risky choice, before they knew it was risky, whereas the impaired patients never showed this anticipatory response. The experimenters concluded that some sort of unconscious biasing in the decision making process and in the strategy choice must have taken place in normals.
Hypnotic studies show plenty of cases where quite complex purposeful behaviour, which is even using information in the form of explicit rules, is nonetheless unconscious. For example, in Sheehan and McConkey (1982) and Spanos (1986), subjects were given an hypnotic suggestion to count but miss the number 4. When requested to count normally, the subjects would miss the 4 and, if asked to comment about what they were doing, they did not report any anomaly in their counting. The subjects were therefore obeying a complex rule, even though they were unaware of it, and also, if asked, denying that they were actually counting in an unusual way. This is an example of a quite complex behaviour, that uses unconscious rule based knowledge in a novel situation.

Of course these studies don't end the research on these topics, especially the hypnotic ones, given the controversial nature on this phenomenon, nonetheless there is sufficient data to require further testing.

For other experimental evidence on implicit planning and related issues, a recently published book by D. M. Wegner (2002), The illusion of conscious will, constitutes an interesting review.

5.3 Further questions on unconscious planning

I hope this chapter shows the importance of a number of questions related to the existence of a capacity for unconscious generation and evaluation of actions, which I have called unconscious planning. This is what is modelled in this thesis, and the very few existing experiments sustain my claim that this is an activity that takes place in the PFC in humans. It is through this basic activity, together with language and culture, that the more familiar process of conscious planning is then based.

I model unconscious planning, as I am going to describe in details in the next chapters, as the capacity of the single brain, to construct expectations about when, where, and what events will occur in the future, depending on which actions are performed. It also involves the learning of these expectations and their organizations in hierarchies, the modulation of relevance of goals depending on the current state of the body (motivation), and the capacity to perform all these activities in parallel. All
these activities are formed only on the basis of reward signals. Some preliminary experiments do suggest that such a biological basic unconscious planning capacity exists in humans, but it has been studied relatively little and many questions remain unanswered.

For example, is the unconscious planning made only as I will model it here, that is to say not symbolic or rule based, or it also might be more complex? Could it be like conscious planning, but “minus” consciousness, indeed explicit, declarative, rule based, but not accessed consciously? Or are these unconscious activities completely unrelated to the conscious one, and working in completely different ways? Or perhaps both kinds of activity, declarative and implicit can take place unconsciously? How are represented implicit contents?

Can we say that all mental phenomena are potentially conscious, indeed with some effort we can become aware of them, we can describe them, feel them, report them to other people and to ourselves? But the range of mental phenomena, conscious or potentially conscious, does not necessarily include all the phenomena having an influence on behaviour and taking place in the brain or the rest of the body. The brain processes information also in exclusively unconscious ways. Dennet is extreme in this position and says “The control of reflexes in man is subconscious, as are the stages of perceptual analysis, and in fact all information processing. We are not aware of the processes at all (as one might, with suitable incisions and mirrors, be aware of one ’s digestive processes)”. Similarly Lashley says, “No activity of mind is ever conscious.”(1966, p.128), where mind here must be read as brain activity.

I wouldn’t be as restrictive as Lashley. I would support the general open research hypothesis that some of the content of the brain processes are conscious, some content could become conscious but is not, and some other cannot. Note that the physical processes in themselves can never become conscious, because they are chemical and electrical processes that have nothing to do with mental processes; it is their algorithmic processing that can lead to consciousness.

Another fundamental question regards the utility of such a conscious/unconscious distinction. Mathis and Mozer 1999, after having equated consciousness to temporally persistent active states and to a clean-up process with well formed outputs, propose
that unconscious brain processes might take place in parallel, whereas conscious processes are serial. The computational utility of this distinction would be given by the fact that, while the parallel processes can be evaluating and processing contradictory options, the conscious processes would only consider single actions that cannot interfere with each other, and so allow agents to execute only one action at a time.

This interpretation fits well with our model of unconscious planning, because it is a parallel process: this leaves space for evaluative conscious processes, based on social inputs, that would make win one of the options over the other based on social values. Plans, or courses of actions through time, can interfere, but consciousness might be a way to avoid this interference, based on social inputs. Consciousness might be a way, biologically determined through evolution, or otherwise possibly just culturally achieved, to give an higher degree of social coherence to a group (Jaynes 1976, Barlow 1987), in conjunction with basic unconscious processes that also are known to be based on social priming (e.g., Bargh 1999). I don't address this aspect of conscious planning in this thesis, much more theoretical and experimental work needs to be done on this matter.

In conclusion, I hope this chapter gives some basis to my claim that planning in humans must be studied first of all as an unconscious process, in order to avoid complex social considerations, that can be made only afterwards. Our unconscious planning is an algorithm implemented as a biological process, and it involves the PFC, the basal ganglia, posterior areas of the cortex, and the cerebellum of a single brain. This idea of a biological algorithm leads us closer to the computational considerations and artificial planning, that I am going to expand in the next chapter.
Chapter 6

Hypothesis 1: planning and action chunks

A man just beginning to learn radio-telegraphic code hears each *dit* and *dah* as a separate chunk. Soon he is able to organize these sounds into letters and then he can deal with the letters as chunks. Then the letters organize themselves as words, which are still larger chunks, and he begins to hear whole phrases.

... In the jargon of communication theory, this process would be called *recoding*. The input is given in a code that contains many chunks with few bits per chunk. The operator recodes the input into another code that contains fewer chunks with more bits per chunk. There are many ways to do this recoding, but probably the simplest is to group the input events, apply a new name to the group, and then remember the new name rather than the original input events.

Miller, G. A. (1956)

In this thesis I emphasize an interpretation of the role of the PFC in behaviour as a planning function. Planning is the capacity of an agent to design goal oriented sequences of actions, without actually executing them, with the scope of evaluating several of these sequences, and possibly execute only the one which receives, given current knowledge, the highest credit in its probability to reach the goal.

A planning function certainly belongs to humans, as seen in a previous chapter, and several researchers have also studied aspects of such capacity in animals (e.g., delayed response tasks, but more in general see Griffin 2001). It is a common practice to view animals’ actions as based on past experiences, but fundamentally oriented towards a more or less distant future, and mostly involved in the struggle to survive and to sustain the delicate equilibria of their bodies in uncertain environments, by predicting and controlling these uncertainties as much as possible. The PFC...
Hypothesis 1: planning and action chunks

constitutes the highest product that evolution has achieved to win this struggle, and a sophisticated planning capacity is its activity.

But planning is not an exclusive capability of animals. In fact, all along the last century the general mechanical process of organizing and evaluating sequences of actions in order to reach a goal has been subject of study also with regard to non biological agents. The idea I want to put forward in the first part of this chapter is that the mathematical or rigorous analyses developed within the framework of how artificial systems might achieve goals of various sort, that is in disciplines known as cybernetics and artificial intelligence, can help in the study of biological systems, and especially PFC activity. With this aim in mind, I am going to review the fundamental approach of these disciplines.

My claim is that the study of the function of the PFC, cybernetics and artificial intelligence are all about the same key computational process, namely planning, and therefore these studies could all benefit each other.

As the problem of how the feet can support the weight of the body is essentially (i.e. beyond the actual and complex biological solution) a problem of mechanical physics and engineering, the problem of how the PFC does what it does is essentially (again, beyond the actual and possibly complex biological solution) a problem of cybernetics and artificial intelligence (independently of what these disciplines have actually historically achieved). It is surprising to note that researchers in cognitive science, that was born on the intention to get inspiration from studies on artificial agents to better understand the living ones, have not, with some notable exceptions (as Miller, Galanter and Pribram 1960, that I am going to describe in details), focussed primarily or explicitly on PFC activity.

In this chapter I illustrate better my claim, and then I concentrate on one formal approach, known as Reinforcement Learning, and its latest developments, as a useful tool to model PFC activity, but I leave open the possibility that other formal approaches in artificial intelligence or in control theory, such as hierarchical control theory, might be even closer in describing the actual planning processes that take place in the living tissues.
6.1 Cybernetics and control

Norbert Wiener defined cybernetics in 1948 (see Cordeschi (2002a) for an introduction to the main concepts) as the study of “control and communication in the animal and the machine”. Its programme involved control theory and information theory, but also neuroscience, philosophy and psychology. The aim was to construct a unified theory of the behaviour of living and artificial agents, both intended as self controlling devices, capable of reaching goals (in the case of the artefacts, the goal is usually set externally by a human).

One key idea of cybernetics was the notion of negative feedback, consisting in an error signal produced by the comparison of the goal state and the current state. The higher is the error, the higher is the need to control and modify the policy to reach the goal. At the time there was also a physiological counterpart of the idea that the body of living systems could control homeostatic variables like temperature or blood pressure.

In 1943 a sort of manifesto of cybernetics was written by Wiener himself, Rosenblueth, who was a physiologist, and Bigelow, who was an engineer. The title of the article was “Behaviour, purpose and teleology”. They classified behaviour into “active”, in which the object itself is the source of the energy in the output, and “passive”, in which all the energy in the output comes from the immediate input. Active behaviour is then divided into “purposeful” or “goal-directed”, and “non-purposeful” or “random”. A system is “purposeful” if it has some specific final condition towards which its activity is directed. Purposeful behaviour can be “with feedback” or “teleological” (from the Greek, telos, which means goal), and “without feedback” or “non-teleological”. Feedback systems are those in which the input is altered by the output in the direction necessary to reduce the discrepancy between the situation so far achieved and the goal situation. Feedback behaviour may be then classified into “extrapolative” or “predictive”, and “non-extrapolative” or “non-

---

1 If instead of goals we talk of rewards, this idea of an error signal is very close to the interpretation of the role of dopamine in the basal ganglia. In PFC I instead interpret the dopaminergic signal as a negative feedback, relative to when, what and where a goal related stimulus (or the goal itself) is expected to occur in the environment. This will be explained also in details in the next chapters.
Hypothesis 1: planning and action chunks

predictive”. In extrapolative behaviour the path of the target is predicted, and the aim is pursued by considering its most probable future position. Such prediction may be of first, second, or higher order. Human behaviour may be distinguished from that of other animals by the use of higher orders of prediction. Given this classification, the authors claimed that it was time for modern science to include in its vocabulary teleological terms like purpose, means, ends. These terms had been disregarded because they referred to future states, that in a causal (one-way) interpretation of the world could not possibly have any influence on a present state. But in purposeful behaviour it is the notion of feedback that can solve the riddle of the future cause. And the attribute of purposeful can be applied not only to biological systems, but also artificial.

Although this cybernetic idea received considerable attention only in the ’40s, it must be said that it was not new, but it was already present in behaviourist psychologist as E. Thorndike and C. L. Hull, (as described in the historical reconstruction of Cordeschi 2002b), of which I will talk again in a next chapter in the restricted context of animal conditioning. Thorndike stated a trial-and-error learning Law of Effect, in which it is the effect that reinforces the correct response among several random actions generated during learning, giving the basis for a psychological feedback mechanism. Hull, as early as in the ’20s, started a “robotic approach”, and his view was that machines could behave like animals, and viceversa organisms’ behaviour could be explained by the same physical laws used to describe machines.

The Cambridge psychologist K. Craik was first to note in the ’40s (Craik 1943, p.52) that Hull’s approach was radically new, compared to previous mechanicism, in his conception that learning models were actual working implementations in the nervous systems, and not generic metaphors. This idea of working learning models is fundamental also in Craik’s view, and he makes it more explicit in the notion of an “internal model” of the world.

The life of this cybernetical research program in its original form was destined to be short. The birth of the more promising approach called artificial intelligence, together with a number of critiques, extinguished cybernetics almost completely between the ’60s and the ’70s. There were two fundamental arguments against it. The first
regarded the idea that the feedback signal needs an existing and explicit goal state with which compare the current state. But organisms, such as humans, have goals that do not exist in reality, but they are the content of wishes or beliefs, and therefore no feedback signal could be derived (Taylor 1966). The second criticism, not unrelated to the first, was that, in order to study teleological behaviour, it was necessary also to have ideas about artificial planning procedures, and not just control. I explain this last point more in detail, given its relevance to our more general topic.

Control (as defined in control theory) in its simpler form is the process of making sure that specific variables defining a system stay as close as possible to some reference (i.e., goal) values. We can say that simple control is an on-line process, because it usually operates through a feedback signal, indicating how much the disturbances bring the values of the variables away from the reference values. More complex control involves the idea that the variables must be regulated according to precise time requirements and variable hierarchies, or, in other words, according to a plan. The complex control procedure is therefore the process of making sure that the difference between what actually happens and a plan is as much as possible close to zero.

Control, therefore, in its simple form, is constantly acting, whereas in a more complex form involves delayed actions and the use of a plan that establishes when the control must be activated. Planning itself is a process that, although possibly oriented towards the realization of its purposes, is nonetheless not necessarily involved with the execution. It is a process that has to do with the generation and evaluation of abstract sequences of actions (the control of variables) to attain complex goals, and it has to do with the anticipation of goal related events (e.g., what, when and where they occur), according to the possible (sequence of) actions that are executed. If a plan has been chosen, its execution becomes a control problem. Control and planning are therefore two necessary and complementary aspects of any complex purposeful behaving system, living or artificial.

Cybernetics, although it claimed its interest in complex behaviour, ended up studying only the control phase of behaviour, that, even if fundamental and necessary, should
be complemented by the planning phase. And planning became the main research interest of artificial intelligence.

6.2 Artificial intelligence and planning

Planning is the core topic of artificial intelligence. It is defined as the study of the algorithmic processes of generation and evaluation of sequences of actions to attain goals (Russell and Norvig 2002).

In early classical planning, formal planning assumed also that the outcome of every action is perfectly predictable. Classical planning consists typically of some kind of logical inference, as in the situation calculus developed by John McCarthy in the '60s, where there are axioms on the consequences of primitive actions and theorems are action sequences, or STRIPS developed at Stanford in the '70s by R. Fikes and N. Nilsson, where an action is defined by its precondition and effects and it is formally an atomic formula as in predicate calculus. There were several descendants of these systems (NOAH, NONLIN, SIPE), but the very limited results in their planning capacities changed in the mid '80s the premises of classical planning to relax the assumption of full environmental predictability. A new generation of robots that should have been able to deal with changing worlds was studied especially at MIT, but the absence of symbolic processing in these machines left them stuck to very simple behaviours, and long term goals were impossible to obtain.

One problem that these studies have, in respect to the main issue here of looking for ideas to be used in a computational model of PFC activity, is that most of these artificial intelligence planning systems use descriptions in the form of first order logic or some subset of it. Therefore states (including goals) are represented by logical sentences, and actions are represented as preconditions and their effects. This seems to be already a quite high level planning capacity. These systems are all inspired by human planning capacities, and some scientists made the appeal to the human way as the centre of their research. I am referring to the extremely influential work, along the second half of the last century, of A. Newell and H. Simon, together or separately, that while studying human performance in problem solving (i.e. generating sequences
Hypothesis 1: planning and action chunks

of goal directed actions) and decision making (i.e., evaluating and choosing these action sequences), they produced also formal mechanism that could achieve similar results. They were centrally concerned with formalizing how people cut problems down to size, and how they apply approximate, heuristic techniques to handle complexity that cannot be handled exactly. One of the main methods they studied is the General Problem Solver, and its heuristics of means-ends analysis.

The huge amount of studies of Newell and Simon could very well be a general theoretical framework for experiments (neuroimaging, for example) on high level PFC function, although none of the authors had an explicit interest in the brain. But these studies seem to me to fall into the category (discussed abundantly in the previous chapter) that interprets planning as a high level process, symbolic, and where all that is needed to know can be expressed in verbal reports (the so called protocol analysis, commonly adopted by Newell and Simon). My main claim, that I support in chapter 5, is that symbols cannot be a starting point in the study of neurobiological algorithms, even though they might arise at a certain level that involves interactions between agents.

An influential work that indeed acknowledges the theories by Newell and Simon, but put more emphasis on the key role of planning in every aspect of behaviour, is Plans and the structure of behaviour, written by two psychologist, G. A. Miller and E. Galanter, and a neuroscientist, K. H. Pribram. The exposition of their basic ideas is very relevant here, because I share their planning and PFC function view in full.

Their book was, together with the early work of Chomsky in Linguistics, considered among the key texts in psychology's shift from behaviourism to cognitivism in the 50s and 60s. It was the first early synthesis of the view that human behaviour could be described in computational terms, thus contributing to the definition of artificial intelligence and the psychological school later known as cognitive science. The authors view "planning" as a pervasive mechanism in behaviour, in the sense that behaviour has always a structure that arises through the execution of plans, which also have a structure. Their definition of a plan is (Miller et al. 1960, page 16):

*any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed.*
Hypothesis 1: planning and action chunks

The basic unit that structures Plans is given in Figure 12:

![Diagram of TOTE unit]

Fig. 12
TOTE: the computational unit of the nervous system in Miller et al. 1960

They call this unit TOTE (Test Operate Test Exit), and it adopts, as explicitly stated by the authors, the "cybernetic hypothesis" that the fundamental building block of the nervous system is the negative feedback. In fact the TOTE pattern is to test the inputs (goal) with the actual state; if there is an incongruity some response must be operated, until the incongruity is eliminated. The operational phase of these units may itself consist of a string of other TOTE units, and so on, until some primitive level is reached, thus constructing a hierarchy. Planning is defined as constructing a list of tests to perform once an Image of the goal has become clear. The sequence of tests is a strategy that leads to the Image.

In their book, the authors extend the TOTE idea to any brain function, from the instincts and motor skills, to memory and language. At the highest levels of the TOTE hierarchy you find the planning function in its more complex form, and in the brain the key role at this level is played by the PFC. The process starts with a decision about what goal should be reached, and then the perceptual patterns that are associated with reaching that goal are activated. Suppose the goal is to pick up a book, then memories of what the book would look like after being picked up must be partially activated. The partially activated goal memory allows a comparison of the current state of the environment (for example, book on the shelf) with the goal state (book in hand). The perceptual patterns underlying the goal state continue to be active until the sensory inputs match the Image (i.e., the sight of a book in hand). So, the frontal lobes...
continually consult the rest of the cortex to see if the current, incoming stimuli are the expected ones so that the next plan can be started.

Although this book is considered to have been highly influential, I have not found it mentioned in many subsequent studies on PFC function. Nonetheless I think the ideas exposed in the book, not only with regard to PFC, but also in respect to other brain functions, constitute a convincing basis to link the brain to behaviour.

That book has two obvious limitations: one is the poverty of experimental evidence to justify the view of the PFC role (the references in the book are all from the 50s backwards, and therefore not many experimental techniques to study the brain were available at the time), and the other is that the formal mathematical aspect are mostly omitted, and only the general TOTE idea is illustrated. The ideas I propose in this thesis can be seen as an attempt to overcome these two limitations, the former by putting together and reinterpreting recent experimental support for such a view of planning, the latter by adopting specific mathematical techniques and constructing computational models.

In the next section I show in details one final kind of artificial planning approach which is highly relevant to us, because its techniques are key to the computational models of PFC activity I develop in the next chapters. This approach is called Reinforcement Learning (RL).

6.3 Reinforcement Learning and planning

RL (for a manual, see Sutton and Barto 1998) is a non classical approach to artificial intelligence. Its central view is that artificial agents can optimize their behaviour by interacting with the environment and trying to maximize a scalar reward signal. Classical artificial intelligence emphasizes instead learning from a teacher or complete knowledge of the world.

6.3.1 The RL framework

The modern formal framework in RL is given by Markov Decision Processes, where:
Hypothesis 1: planning and action chunks

- the agent and the environment interact in a sequence of discrete time steps \( t = 0, 1, 2, 3, \ldots \);
- on each time step \( t \), the agent sees the environment to be in a particular state \( s_t \) and it selects action \( a_t \);
- the response of the environment to the action is to stochastically transit to a new state \( s_{t+1} \) and emit a reward signal \( r \in \mathbb{R} \).

Most of the RL problems are cases in which the state and action spaces are finite. Central to the RL theory are the concepts of policy, which is the mapping from states to probabilities of taking each action, and of value function of a state \( s \) at time \( t \) following policy \( \pi \):

\[
V^\pi s(t) = E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots]
\]

where the discount factor \( \gamma \) specifies how much in the future rewards should be taken into account. The value of a state is the expectation of future cumulative rewards. What is the optimal policy for the system? The one that maximizes this expectation, and it is given by solving the Bellman equation, at the core of all the RL techniques, as well as dynamic programming (Bellman 1957):

\[
V^*(s) = \max_a [r(s,a) + \gamma V^\pi(s'(s,a))]
\]

where \( r(s,a) \) is the reward that the agent gets if it takes action \( a \) when in state \( s \), and \( s'(s,a) \) is the state of the environment after action \( a \) has been taken at state \( s \) following policy \( \pi \). The Bellman equation states that when evaluating an action to be made at time \( t \) and state \( s \), both the immediate reward \( r(s,a) \) and the future cumulative discounted reward \( \gamma V(s'(s,a)) \) must be considered. The idea that the estimate of the value of a state is based on the values of successor states is called bootstrapping, and it is an interesting feature of many RL methods.
6.3.2 Temporal Difference learning

There are different methods that learn value function and selection of the optimal actions. For us the most interesting, given its simplicity and its biological plausibility (as we’ll see in the next chapter) is Temporal Difference (TD) learning. It states that the error in the selection of action \( a \) taken at state \( s \) is given by:

\[
\delta(s, a) = r(s, a) + \gamma V^*(s'(s, a)) - V(s)
\]

which can also be written in function of time, thus giving better the sense of why it is called TD:

\[
\delta(t) = r(t) + \gamma V^*(t+1) - V^*(t)
\]

This \( \delta \) can be interpreted as the error in the prediction of the immediate and future discounted rewards. From this formula it can be seen that, at the early stage of the learning of an association between a state and a delayed reward, this delta is positive immediately after the delivery of the reward, because \( r(t) \) is positive and the several \( V(t) \) are still zero. As learning progresses, and the reward is more and more predicted, the positive value of \( \delta \) is more shifted backwards in time, towards the time occurrence of the state predictive of the reward. Moreover if, after learning, the reward were suddenly omitted, the value of \( \delta \) would become, in the moment the reward was expected, negative, because \( r(t) \) would be zero, \( V(t+1) \) would also be zero, but \( V(t) \) would be positive, because a reward was expected.

There are several ways to select actions during initial stages of interaction with the environment, when not much learning has been acquired, and instead exploration is required. As learning improves, the exploratory behaviour should be reduced, and less randomness (noise) should be allowed. One way is to make the probability that a certain action is chosen at time \( t \) to be:
Hypothesis 1: planning and action chunks

Pr ob\( (a(t) = a_i) = \frac{\exp[\beta \delta(s(t), a_i)]}{\sum_j \exp[\beta \delta(s(t), a_j)]} \)

where \( \beta \) is called the inverse temperature and controls the randomness in the choice of the action. Another similar method is to make the randomness be inversely proportional to the number of rewards, following a sigmoidal function. The estimate of the value function is given by:

\[ V^\pi(s(t)) = V^\pi(s(t)) + \alpha \delta(s(t), a(t)) \]

where \( \alpha \) is the learning rate, that is how much of the error signal \( \alpha \) we want to have an effect on learning.

6.3.3 Policy iteration

These simple RL definitions constitute the theoretical basis for methods allowing an artificial system to learn which is the best policy to be adopted in a changing environment, not only when the rewards are given after each single action, but also when they are only given after correct sequences of actions. One important method, which interests us for its biological affinities, is called policy iteration. This method consists of two components: the critic, that uses TD learning to estimate the total future discounted reward that is expected when following the current policy, and the actor, which maintains and improves the policy. The choice of the actor of which single action to take is based not on the immediate reward, but on the estimate of the future discounted rewards as provided by the critic.

To account for experiments in which animals and multiple environmental states, i.e. stimuli, are involved, when these actor/critic systems are used, two vectors must be introduced: the stimuli, represented by a vector of binary variables \( u \), with each component representing the presence or absence of the relative stimulus, and the weights \( w \), such that the expected reward is given the dot product:
Hypothesis 1: planning and action chunks

\[ v = w \cdot u \]

and the weights are updated by the following learning rule (gradient descent):

\[ w = w + \alpha \Delta u \]

6.3.4 RL and planning

What does have RL have to say with regard to planning, the central process of interest here? In this case a key definition is the *model* of the environment, which is used to simulate interactions between the agent and the world. This simulation is then processed by RL methods just as those seen previously. In this way the policy gradually improves over time. A key aspect of this simulation it that it can abstract over time. In fact in RL planning is defined as the process of achieving a goal state by choosing among different courses of actions that span over a different range of time scales, for some of which we don't know *apriori* when they terminate.

For example, imagine that you are in your new office at your desk in a town you don't know much, and you want to catch a bus to go to a library in another part of the city, where you know there is a book you are looking for. Here the goal is constituted by the book that you want to have, and the reward will be obtained when you actually have the book. But before you can reach it, a number of different sets of actions or courses of actions (sort of subroutines), acting at different time scales, must be performed: leave the desk, take your jacket, lock the office, leave the building, find the closest bus stop, check what's the right bus to take, wait until it arrives, et cetera.

So, the very high level action is to go to the library and get the book, but, to achieve that, a plan is needed, and you have to explore in the hierarchy of actions you have to perform, from higher level actions like "taking the bus", to lower level actions like "locking the door with your key", or even "use this and this muscle to turn the key". For some of these courses of actions you don't know exactly how much time they will
take (when the bus will actually arrive), but in our planning we have two fundamental expectations:

- an expectation of roughly how much time each course of action will take (half an hour of waiting at the bus stop);
- an expectation of what will be the state after each course of actions is finished.

In our example, when thinking to you taking the bus, you don't need to know all the details of how actually you will get on it, what your movements will be, et cetera: you only need to know that the bus will probably arrive within an hour, and when this happens, you will be on the bus.

6.3.5 Temporal abstraction and options

We also want to learn from the actual execution of this plan, so that the next time we are in the office and we want to go to the same library, we want to control and organize our actions efficiently (not to wait too much at the bus stop, or get off at the stop that is really the closest to the library).

The analysis of decision processes acting at varying time scales, has been explored in artificial intelligence and control theory, but recently also, in the last couple of years, by RL. RL presents a general mathematical framework for prediction, control and learning at multiple time scale, and called Temporal Abstraction (TA). It is on TA that I will concentrate, because at it deals with the basic simple mechanisms for planning we have just described. Now I summarize some of the core definitions regarding TA, but for a full exposition of this growing area of RL see Precup 2000.

TA refers to the use of high level actions, whose execution can take varying amounts of time steps which cannot be determined apriori, as opposed to low level actions, which are assumed to take a uniform amount of time. The intuitive idea behind the use of TA in RL is very simple: if an agent has to take a decision on how to reach a goal state, it is allowed to choose among different options, that is to say to predict
Hypothesis 1: planning and action chunks

what happens (i.e. what will be the final state) if it follows whole courses of actions that can be temporally extended, stochastic, and contingent on previous events. An option is a closed-loop policy, which is initiated, it decides which actions to take for some time, and then terminates. When an option terminates, the agent can choose to start others, until it reaches its goal state. In this hierarchical and temporally abstract decision making process, the agent achieves faster planning than if using just one level of time scales, simply because each option, when considered at its higher level, takes just one time step to be evaluated, even if its actual execution might take several steps.

These ideas have been formalized within the RL mathematical framework. Options are defined by three elements:

- the set of states \( I \) in which the option applies;
- a decision rule \( \mu \) which specifies what actions are executed by the option;
- a completion function \( \beta \) which specifies the probability of completing the option on every time step.

An option is available in state \( s \) if and only if \( s \) is in \( I \). If the option is started, then the actions are selected according to \( \mu \), until it terminates stochastically according to \( \beta \), then a new option is chosen, also according to \( \mu \), and so on.

Note that also single actions can be treated as options: they are initiated in a state, take control for one time step, and then end. Therefore, \( \mu \) can be generalized to choose among options (including single actions). The Bellmann equation given before requires that the policy’s decision probability at time \( t \) should be a function only of the current state \( s \). This type of policy is called Markov. When dealing with the policy \( \mu \) of options we allow the probabilities of selecting a sub-option to depend on all the states and actions from time \( t_0 \), when the action began, up to the current time \( t \). This type of policies are called semi-Markov, and we are in the presence of a Semi Markov Decision Process.
Also the completion function $\beta$ can be semi-Markov. This allows several ways of ending an option, for example after a fixed number of steps, after a fixed amount of time, or after the occurrence of a critical state, which might be a desired subgoal. Two options $a$ and $b$ can also be composed to produce a new option, $ab$, that first follows $a$ until it terminates, and then follows $b$, whose termination also ends $ab$. Given this fundamental definition of option, now we must allow the agent to predict what happens when one or another option is executed: in this way planning is allowed over temporally abstract courses of actions. To have this, for any option $o$, be $e(o, s, t)$ be the event of $o$ being started in state $s$ at time $t$. Then the predicted reward is:

$$r^*_o = E[r(t+1) + \gamma r(t+2) + ... + \gamma^{(k-1)}r(t+k) | e(o, s, t)]$$

where $t+k$ is the random time of termination of $o$. The prediction relative probability of terminating $o$ in state $s'$ is:

$$p^o_{s'} = \sum_{k=1}^{\infty} p(s', k)\gamma^{k}, \forall s' \in S$$

where $S$ is the set of all the states, and $p(s', k)$ is the probability that $o$ terminates in $s'$ after $k$ time steps. So note that $p^o_{s'}$ is a combination of the likelihood that $s'$ is the terminal state for $o$ together with a measure of how delayed that outcome will be, relative to $\gamma$. These two predictions, one for the reward and the other for the final state, are called a model, and in particular, given the capacity to plan abstracting from time, a multi-time model of an option.

The value function for a state can then be rewritten to include options $\mu$ as:

$$V^\mu(s(t)) = E[r(t+1) + ... + \gamma^{k-1}r(t+k) + \gamma^kV^\mu(s(t+k)) | e(\mu, s, t)]$$

where $k$ is the duration of the first option selected by $\mu$. It can be shown that the Bellman equation can be rewritten as:
Hypothesis 1: planning and action chunks

\[ V^*_o(s) = \max_{o \in O} E[r + \gamma^k V^*_o(s') | e(o, s)] \]

where \( O \) is the set of all the available options, \( r \) is the cumulative discounted reward received along the way, \( s' \) is the next state.

6.3.6 Learning and options

Methods used to learn value functions and selection of the optimal actions, as the TD learning, can now be generalized to learn value functions over models of options. Different ways to learn can be configured: some methods (semi-Markov) require an option to be executed until termination, other, more biologically appealing, that learn about an option from small fragments of experience consistent with that option, even if the option itself is not executed. These last methods are called intra-option, and they are a kind of learning takes place off-line, i.e. the consequences are learned while potentially behaving according to another option. In these methods, options not executed and not relevant to the reward that the system is trying to maximize, can nonetheless be improved for a future use (if when searching for your keys, you also see that you left your cup on the desk, you might need this information when you need the cup).

The TD learning version for off-policy intra-option value learning of multi-time models of option, \( r^o_s \) and \( p^o_s \), given the knowledge of the option (its \( \mu, \beta \), and \( I \)) is relevant to our model. The simpler way to learn the model of an option is without the intra-option possibility, to execute the option to termination many times in each state \( s \), record the next states \( s' \), the cumulative discounted rewards \( r \), and elapsed times \( k \). These outcomes can then be averaged to approximate the expected value for \( r^o_s \) and \( p^o_s \), given before. For example, an incremental learning rule for this could update its estimate \( \hat{r}^o_s \) and \( \hat{p}^o_s \forall x \in S \), after each execution of \( o \) in state \( s \), by:
Hypothesis 1: planning and action chunks

\[
\hat{r}_s^o = \hat{r}_s^o + \alpha [r - \hat{r}_s^o]
\]

\[
\hat{p}_{as}^o = \hat{p}_{as}^o + \alpha [\gamma \zeta_{s's'} - \hat{p}_{as}^o]
\]

where \( \zeta_{s's'} \) is an identity indicator, equal to 1 if \( x=s' \) or 0 otherwise; \( \alpha \) is a step-size parameter, which can be constant or can depend on the state, option and time.

Now let's consider the intra-option method. The idea is to use the Bellman equation, or a version of it, as we did for the simple actions without options. Indeed, the correct model of a Markov option \( O=(I, \pi, \beta) \) (that is the version we see here, but it can also be extended to semi-Markov options) is related to itself by:

\[
\hat{r}_s^o = \sum_{a \in A_s} \pi(s, a) E[r + \gamma 1 - \beta(s')]r_{s'}^o = \sum_{a \in A_s} \pi(s, a)[r_{s'}^o + \sum_{s'} p_{as}s' (1 - \beta(s'))r_{s'}^o]
\]

where \( r \) and \( s' \) are respectively the reward and next state given that action \( a \) is taken in state \( s \) from \( A_s \), the set of available actions. Moreover:

\[
p_{ss'}^o = \sum_{a \in A_s} \pi_a[1 - \beta(s')]p_{s's'} + \beta(s')\zeta_{s's'} = \sum_{a \in A_s} \pi(s, a) \sum_{s'} p_{as}s' (1 - \beta(s'))p_{s's'} + \beta(s')\zeta_{s's'} \quad \forall s \in S
\]

These Bellman equations can be translated into the TD update rule by considering that action \( a(t) \) is taken in \( s(t) \) and that the way it is selected is consistent with \( o=(I, \pi, \beta) \).

So we have:

\[
\hat{r}_{s(t)}^o = \hat{r}_{s(t)}^o + \alpha [r(t+1) + \gamma (1 - \beta(s(t+1)))\hat{r}_{s(t+1)}^o - \hat{r}_{s(t)}^o]
\]

\[
\hat{p}_{s(t)x}^o = \hat{p}_{s(t)x}^o + \alpha [\gamma (1 - \beta(s(t+1)))\hat{p}_{s(t+1)x}^o + \gamma \beta(s(t+1))\zeta_{s(t+1)x} - \hat{p}_{s(t)x}^o]
\]
Hypothesis 1: planning and action chunks

This method is called the one-step intra-option model learning, and it is the simplest intra-option model-learning method that is possible to conceive, but it is possible to add to it standard tricks for off-policy methods in RL. In this formulas the equivalent of the $\delta$ at the level of the primitive actions, is divided in an error for the prediction of the reward:

$$\delta_{\text{reward}} = r(t+1) + \gamma(1 - \beta(s(t+1)))\hat{p}^o_{s(t+1)} - \hat{p}^o_{s(t)}$$

and an error in the prediction of the probability of resulting in a certain state when the option ends:

$$\delta_{\text{state}} = \gamma(1 - \beta(s(t+1)))\hat{p}^o_{s(t+1)x} + \gamma\beta(s(t+1))\zeta_{s(t+1)x} - \hat{p}^o_{s(t)x}$$

6.3.7 Synchronous Value Iteration

Let's go back to planning in RL. Planning, as mentioned, is defined as the computational process that takes a model of the environment as input and produces or improves a policy for interacting with the environment by searching through the state space. The computational process in itself consists in simulating experience, based on the current model, and in computing the value function over the states, in order to find to best path from the current state to the goal state. Planning is therefore a learning process over a simulated experience. Real experience can then be used by the agent to improve the model and to improve the value function and policy.

Several techniques can be developed according to this definition of planning, but we will concentrate on one in particular, for its simplicity. It is called Synchronous Value Iteration. This technique has been known since the 50s, and was developed by R. Bellman (1957), but it is commonly used also in modern RL approaches. It consists in a parallel step by step evaluation process of the states of the environment, that starts from the goal state, until it reaches the current state of the agent. The best current
Hypothesis 1: planning and action chunks

Policy can then be easily chosen by the agent by simply picking up actions that in each time step lead to the higher evaluation (for more details, see Precup 2000).

If we allow options and TA, a RL agent can simulate and evaluate not only what is the state and the reward after a primitive action is taken, but also which is the state and the reward after an entire option, acting at varying time scales, is taken. In the higher level process of planning, options take just one time step, even if their actual execution might take several steps: this speeds up the evaluation process considerably.

6.4 RL simulation using TA: a motor task

The purpose of the following simulation is to test the effectiveness of options in reducing planning time.

The toy problem under examination regards a motor task. An agent is given several goals consisting of reaching different locations in a discretized ring shaped world with \( n \) states (1, 2, 3, .... \( n \)) as depicted in figure. The units are binary, and they can represent single states.

![Fig. 13](image-url)

A ring shaped grid world. An agent (yellow) has to reach a goal location (red).
In each trial the agent (symbolized by the yellow circle in the figure) has to reach a goal location (red rectangle). Its job is to plan, given current knowledge of the environment, the best sequence of actions (sequences of move Left and move Right) to reach its target in the shortest number of moves (i.e., move in one direction and not the other, because it’s the shortest path; but the agent does not have a notion of direction, so it might alternate Left and Right randomly).

A basic assumption of the simulation is that the system takes one time step to go to a next state of the environment, both in the actual execution and in the simulative planning process, but for the planning process it can simulate several actions in parallel. Another simplifying assumption is that the actions are not stochastic, but deterministic.

Initially the agent is constructed so that it knows, and therefore plans, only relative to the one step-actions Left or Right (defined as option 1 and 2). If it moves Left in state \( m \) (where \( m \leq n \)), it can predict it will go to state \( m-1 \) (except if \( m=1 \), then the next state is \( n \)); similarly, if it moves Right in the state \( m \), it can predict it goes into state \( m+1 \) (except if \( m=n \), then the new state is \( 1 \)).

In each trial a reward is randomly put in one of the ring locations. The agent is given information on which is the rewarding state, and its job is to plan the best sequence of actions to reach it, given currently available internal models. Planning comes always before any action, and it is achieved with SVI, i.e. successive approximations of values expectations for each state of the ring, as described in subsection 6.3.7. A value of 1 for a specific state indicates which is the goal. It is also used information on which is the initial state, because the planning stops until the agent reaches (in the simulative planning process) the current state starting from the goal state.

The chunking process is obtained in this way. When the reward/goal is found, the agent constructs a new internal model relative to the specific sequence of actions it took to reach it from the start state (if it didn’t have the internal model already), and a new relative option is added to the available repertoire. Note this is quite dissimilar from the main simulations in Precup (2000), where the options are not generated by the models, but already given to them and fixed.
Hypothesis 1: planning and action chunks

Several multi-time internal models start to be constructed in my model. As several of these new internal models are available, the planning process becomes faster and faster compared to the initial stages, because these longer range models don’t make the agent waste planning time steps in re-evaluating intra-model actions.

To check the computational advantage in the planning process, I did the simulation on a ring with 50 states with and without the chunking mechanism, i.e. only in the former version the agent is allowed to learn the internal models. In the plot I compare performance of the two versions in the following way: I first generate a sequence of 40 tests in the form a start/goal pairs. These tests are of varying complexity, namely the distance from the start location to the goal can be as low as zero (the agent is already in the rewarding location) to 25 (which is the highest distance in a 50 state ring). I then run the simulation on this fixed sequence, both with (blue) and without (red) the chunking mechanism. What I report here is the number of steps it takes to the evaluative planning process to choose the best policy.

![Graph showing comparison of performance with and without chunking mechanism](image)

**Fig. 14**

A comparison of performance in the planning process with and without the chunking mechanism.
As it can be seen, the chunking process improves planning performance considerably. At the beginning, up to around test 5, the two performances are similar, namely the planning process takes the same times steps to evaluate states. But later on, the chunking agent (blue) takes always less and less time steps compared to the non chunking one (red), except for those tasks where the distance between the start and goal locations are short (i.e., one or two simple actions away).

Note that the actual difficulty (i.e., start/end distance) of each test in the fixed sequence is equal to the number of planning steps it takes to the non chunking mechanism to reach a solution.

In the last 10 tests the chunking agent plans in just one or two time steps the entire sequence of correct action to reach goal location, whereas the non chunking agent still needs to evaluate all the single actions, and so it takes up to 20 time steps in one case (test 38).

6.5 The biology of action chunking

From these considerations on artificial systems, it should be clear that chunking, corresponding in RL to the construction of options, is an essential component of any efficient planning device. Therefore, let’s transfer this idea of chucking actions into the biological domain, and let’s see if there has been any study concerning the production of such faculty in a nervous system.

The idea of action chunking in the brain has been actually suggested by A. M. Graybiel, and the best general description can be found in Graybiel 1998. The basic argument is that the dorsolateral striatum in the basal ganglia is involved, during the course of Stimulus-Reinforcement learning, in a recoding consisting in chunking the representations of motor and cognitive action sequences, so that they can be treated by the nervous system as a unit. This argument is not in contradiction with other existing interpretations of the basal ganglia as a selector of actions (e.g., Gurney et al 2001), but it is adding the hypothesis that the specific role of the dorsolateral striatum is to form chunked representations of actions, thus contributing to the progressive construction of an hierarchies of actions.
6.5.1 The formation of action chunks

This idea of action chunking is used by Graybiel to explain the acquisition and the structuring of action repertoires, i.e. collections of habits, but my proposal is to extend her use one step further, namely for (implicit) planning. There can be several sequences of actions that can be chosen to reach a novel goal, and we need a process to select which of these is the most convenient depending on the context, and therefore an implicit planning process. The organization of habits in action chunks allows an animal to simulate and evaluate implicitly the optimal collection of habits that must be performed in order to reach the novel target.

Graybiel proposes that the dorsolateral striatum in the basal ganglia is the place where there is a storage of states in which options start and end. My proposal is that this organization facilitates considerably, through abstraction, the evaluative computational planning process; at the same time, the dorsolateral PFC plays the role of learning new combinations of habits taken from the action repertoires, thus allowing the construction of new internal models.

6.5.2 Experimental evidence for action chunking

In Jog et al. 1999 study, rats had to learn a procedural task, consisting in running in a T-maze, composed of a start position with a gate opening, an instruction point where one of two tones is emitted, a turning point where the rat has to turn left or right, and two final goal locations, with only one having the reward, as consistently indicated by the two tones. Cellular activity was recorded in the dorsolateral striatum of four of these rats. A general re-shaping of the firings took place as skill of the maze was acquired. A reorganization of the activity was observed: from distributed along all the phases, to mostly concentrated only at the beginning and end of the procedure. This pattern is interpreted by Graybiel as consistent with the idea that the striatum constructs an action template for the entire procedure, and the start/end activity becomes as a sort of label for the action chunk. The basal ganglia recode cortical inputs into a chunked form, namely as a unitary sequence of actions.
Hypothesis 1: planning and action chunks

In an experiment reported in Graybiel 1998 (page 129), two monkeys were trained on a three button press task, and one animal was lesioned unilaterally to the nigrostriatal system before the training, the other was lesioned after. Learning was showed to take place only if the striatum was intact on the side contralateral to the tested arm. But one particularly interesting finding was evident when the experimenters delivered a reward one button press earlier than expected, i.e. before the last button press. Both monkeys, with the arm contralateral to the intact striatum, continued to press the last button (following their newly acquired habit), even though they had received the reward already. The monkey with the posttraining lesion also made these perseverative errors. The other monkey with the pretraining lesion, did not displayed these errors in the arm contralateral to the lesion. This is interpreted by the authors as a deficit, due to disruption of the striatal system and the neurotransmitters involved (dopamine, in particular), in the formation of a habit, or macro, or performance unit, or action chunk. Studies on the essential role played by the striatum and its interaction with the PFC during skill acquisition in humans can be found in Poldrack et al. (1999), where are reported fMRI results on subjects who learned probabilistic classification, a cognitive task.

It must be noted that the specific location of chunked representations in the dorsolateral part of the striatum appears to be necessary (as also replicated in a different laboratory than Graybiel, see Yin et al. 2004), whereas neighbour areas do not show evidence for such code (Daw et al. 2002).

The recoding taking place in the basal ganglia through the different stages of learning is just one aspect of the plasticity present in the whole brain. Neuronal activity in the striatum depends on several excitatory afferences coming from other brain areas, mostly the cortex, and changes in this basal ganglia structure must be understood in conjunction with parallel activity changes taking place in these other regions. The most relevant of these changes happens in the cortex, where a major shift of the activity from the beginning to the end during motor skill learning takes place from PFC to posterior cortex. Another well known shift takes place in the striatum itself, where activity moves from rostral to caudal (see Miyachi 1997).
Another basic theme in these studies is that the basal ganglia have this recoding role not only for motor action, but also for "cognitive patterns" (Graybiel 1997), that is to say habits of thought. The basal ganglia form associations among PFC inputs, therefore they automatizes sequences of actions, and leave free the computational space in the cortex to evaluate and attempt novel sequences of actions to reach new goals.

One issue to be investigated regards the role of dopamine in the chunk formation. Previous modelling work on the basal ganglia shows how dopamine in the striatum might have the effect of signalling an error in the prediction of rewards, so that neural activity shifts entirely from the time of occurrence of the reward (usually at the end of the trial) to the time of occurrence of the first predictive stimulus (at the start of the trial). But in the chunking process the neural dynamics follow different patterns, because at the end of learning many neurons in the striatum are more active at the start of a trial, but also at goal reaching, i.e. at the end.

One hypothesis to test is that, while the shift of the activity to the start of the trial is explained by dopaminergic effect from the substantia nigra pars compacta, the activity at the end of the trial might instead be due to the connections coming from dorsolateral PFC and its coding of the goal related expectations. Anatomical studies and several available experimental parameters on the reorganization of the neuronal activity in the striatum, such as the number of units active, the changes in individual activity, and timing considerations, set precise constraints which could possibly falsify the hypothesis. In this case, further ideas can be investigated, for example the modulatory role in striatum of neurotransmitters other than dopamine.

### 6.5.3 The selection of action chunks

If a repertoire of options is available, than there is the problem of selecting which of these options is better to follow in order to reach a goal. This process is quite complex and articulated, and I view it as performed by three different brain areas: the striatum itself in the basal ganglia, the anterior cingulate cortex and the orbitofrontal cortex.
In my view, the first selective processes take place at the basal ganglia stage, where label of options are represented. A simple competitive mechanism between the different chunks can make more active only a limited number of them, thus pushing the activity at the cortical level towards certain pattern of activity over others (Djurfeldt et al. 2001). A similar interpretation of the role of the basal ganglia, but exclusively in a motor domain, can be found in Mink 1996. I extend this role also to cognitive programs.

After this low level process, higher level option selection takes place at the cortical level. Several studies are starting to accumulate evidence on how the cortex might actually achieve such a complex function of selecting sequences of behaviours (see Krawczyk 2002 for an extensive review). The two major candidate for such a function are the orbitofrontal cortex and the anterior cingulate cortex. Studies on the neural basis of human decision making suffer from the prejudice that planning is exclusively a conscious process, so that there are only very few experiments (Damasio1997, Berns 1997) that test an unconscious version of this mechanism. It is therefore quite difficult to make convincing statements of this level of the option selection. Nonetheless it is quite clear from several studies, mostly on humans using neuroimaging techniques (see MacDonald et al. 2000 for a review), how the anterior cingulate cortex is involved in evaluative processes indicating when one sequence of actions needs to be more strongly engaged over another. And it is also clear that there is a key involvement of neurons in the orbitofrontal cortex of monkeys in processing motivational aspects of goal related objects, so that neurons can determine the probability of relevance of one sequence of actions over another.

In this respect a key study is reported in Tremblay and Schultz 1999. In this study cellular recordings were made in the orbitofrontal cortex of Macaca fascicularis monkeys, while performing a delayed response task similar to Watanabe 1996. A picture was presented briefly in a left or right location of the visual field, to indicate the target of an arm movement, and its visual features predicted specifically which of two rewards would be delivered for correct performance. A subsequent trigger stimulus provoked the movement to the remembered target. The reward was delivered after a brief delay. The main results are that specific orbitofrontal neurons in
dorsolateral prefrontal seem to discriminate well between different rewards, but here many discriminations seem to be based on the relative preference for different rewards. In fact, prior to the described task the animals are tested in a separated choice behavioural experiment, in which the monkeys show their individual preferences over three different food rewards, namely raisins (A, most preferred), small apple morsels (B, intermediately preferred) and sugar honey cereal (C, least preferred). In the delayed response task, in each session only two kinds of rewards are presented, so that there is always a preferred reward over the other (A>C, A>B, B>C). The activity of some of the cells does not appear to code the fixed identity of the rewards, but instead seems to represent the relative motivational value of one reward in respect to another, as expressed by the individual preference. Orbitofrontal neurons can be more or less activated by one reward, depending on which alternative reward is available. This mechanism is an indicator that a process similar to the construction of a RL evaluation function and based on internal needs or preference is taking place.

With regard to the role of anterior cingulate cortex in the selection process, the interpretation that has been put forward in MacDonald 2000, for example, and other works from the same authors, is that this part of the cortex is involved in conflict monitoring. The idea is that when executing a task, human subjects might face the problem of executing two or more groups of actions in alternation as set by the context, and it has been shown that the anterior cingulate area is more strongly engaged when the number of competing behaviours increases.

6.6 RL simulation using TA: a cognitive task

The purpose of the following simulation is at the same time to test the effectiveness of options in reducing planning time, and to study the performance in a complex and cognitive domain. The focus of this simulation in respect to the general framework illustrated in Fig 10 of chapter 4 is on the basal ganglia as a chunk labelling module, on the premotor and motor module, as storing the actual actions to be executed once an action chunk is triggered, the orbitofrontal cortex as a goal evaluator, parietal and temporal cortex module, that gives the inputs.
Hypothesis 1: planning and action chunks

There is a fundamental problem when attempting to make a computational model of implicit planning and chunking in cognitive situations, and therefore in finding data with which to compare performance between the simulation and real observations. It is related to what I have abundantly discussed in chapter 5, namely that there are not cognitive tests of the unconscious planning capacity, but only of conscious planning, where subjects are verbally instructed on what to do. Therefore, unless it is modelled a part that also deals with verbal instructions, the simulation won’t have much significance.

One type of experiment that would suit my requirement would be cognitive planning tests that are performed on humans as if the setting of the experiment were for monkeys, i.e. the teaching signals are just rewards or punishments (or omission of rewards). Again, these experiments I have unfortunately not found. This model focuses on the implicit aspects of the performance, that, although it is acknowledged that they are present in the execution of basically every kind of task (Broadbent and Berry 1988), I have not found a study relative to the specific test I use here. Therefore I am not going to compare the chunking and planning performance of my simulated model with any of the data available on humans performing this cognitive test. I will instead propose an experiment that can produce data which can be meaningfully compared with the model’s results.

The test that I use is the Tower of Hanoi. It is a game used in cognitive psychology to study explicit planning capacities in humans. There is a small number of disks, typically no less than three, or no more than five, of different size and three pegs. Two example configurations are showed in Fig.15 for the 5 disks version. Any disk can be moved to any empty peg or onto any other disk of bigger size. Given any starting configuration, the task is to reach a goal configuration in the shortest number of moves.
A similar task, the Tower of London, has been used for simulation purposes in Dehane and Changeaux 1997. It is similar to the Hanoi one, because it consists of moving three (or more) coloured beads, mounted on vertical pegs of unequal length, from an initial position to a prespecified goal. The main difference between the two tests is that the London version doesn't have rule based priority ordering (i.e., any bead can go on any other bead). A key difference with my model is that these authors didn't not incorporate any consideration about the chunking process, and, more importantly, their simulation, which has nothing to say about language or verbal instructions, doesn't consider the possibility of an implicit planning capacity. Therefore they model a cognitive test that is known to be instructed verbally, and solved in verbal terms (as is clear from verbal reports), but surprisingly language does not play any part in their simulation.

In my simulation a basic assumption is that the system takes one time step to go to a next state of the environment, both in the actual execution and in the simulative planning process, but for the planning process it can simulate several actions in parallel. Another simplifying assumption is that the actions are not stochastic, but deterministic. The input units are binary, and each represents a different state of the environment, that is to say a different combination of disks on the pegs.
In an initial training phase, the agent doesn’t have any information on which are the allowed moves of the game. Basic shifts of the disks’ positions from one peg to another are randomly attempted, and if they are allowed by the rules of the game, a reward is given (by the simulated experimenter).

The agent therefore learns what are the consequences of moving single disks, in the form of specific actions that lead from one state of the environment to another (basic internal models), and also with an attached predicted reward for each action. No real process of abstraction takes place, and so it’s not the rules of the game that are learned (as a language based system would probably do), but stimulus-response-prediction links, or, in other words, (current_state)→(move)→(next_state & next_reward).

Once all basic possible moves are covered, the system goes into a second testing phase. Random pairs of starting states and goal states are generated. A value of 1 for a specific state indicates which is the goal.

The job of the agent is to plan the best sequence of actions to reach the goal, given currently available internal models. Planning comes always before any action, and it is achieved with SVI, i.e. progressive evaluations of the reward value starting from the goal state, and then progressively of the adjacent states from which there are action that can be used to reach the goal in one step, and then in two steps, and so on, until the current state is reached. At that point the agent chooses the shortest path (i.e., involving the least number of moves) that leads to the goal. The action is actually executed, and, after the goal state is actually reached, a new option is generated, that includes the recent starting state, the goal state, and the sequence of moves just chosen and taken to reach the goal. In this way several multi-time internal models start to be constructed.

This option generation, which is also present in the previously described motor task model (section 6.4), is instead not present in Precup (2000) main simulations, where the options are fixed and given from the programmer.

As this testing phase progresses, several complex tasks are given to the agent, namely sessions and rewards that involve a sequence of moves to reach goals. Schematically the program goes as follows.
I run the simulation of the Tower of Hanoi with and without the chunking mechanism, i.e. only in the former version the agent is allowed to learn and construct internal models and chunks. In the first plot (Fig. 17a) I compare performance of the two versions for the 4 disks version in the following way. I first generate one single random sequence of 60 tests in the form a start/goal pairs. These tests are of varying complexity, namely the minimal distance from the start location to the goal, in terms of primitive moves that can be made, ranges from zero (when start state and goal state are coincident), to 15 (in the case of start equal to a tower with all the five disks, and a goal configuration with the full tower, but on another peg). An higher number of moves compared to the minimal solution can be executed if the system does unnecessary moves that need to be corrected.

I run the simulation on this fixed sequence, both with (performance represented in blue in the figure) and without (performance represented in red) the chunking mechanism. Performance is represented as the number of steps it takes to the evaluative planning process to choose the best policy.

As it can be seen, the chunking process improves performance considerably. At the beginning, up to around test 7, the two performances are similar, namely the planning process takes the same times steps to evaluate states. But later on, the chunking agent takes always less and less time steps compared to the non chunking one, except for those tasks where the distance between the start and goal locations are short (i.e., one or two simple actions away). Note that the actual difficulty (i.e., start/end distance) of each test in the fixed sequence is equal to the number of planning steps it takes to the non chunking mechanism to reach a solution.
Hypothesis 1: planning and action chunks

At the end of the trial, the chunking agent plans in just one or two time steps the entire sequence of correct action to reach goal location, whereas the non chunking agent still needs to evaluate all the single actions. For example it takes up to 24 time steps in test 36, whereas the chunking takes only 5 steps.

In figure 17b I generalize what is shown in figure 17a, in that, instead of showing the performance of the model on just one single random sequence of 60 trials, I generate 70 different sequences of trials, run the 4 disks model with and without the chunking capacity, and then take the mean evaluative time steps for each position in the sequence (i.e mean for all first trials, mean for all second trials, and so on up to the 60th trial).

![Graph](image)

**Fig. 17a**

A comparison in the 4 disks Tower of Hanoi simulation between a chunking and a non chunking agent on one random sequence of 60 different start/end pairs.
Hypothesis 1: planning and action chunks

It can be seen that the average test without chunks takes between 8 and 10 time step to be solved by the model, whereas with the chunking mechanism the diagram show a clear learning curve.

The design of an experiment with which we could compare the results of the model would be defined exactly as the program itself. In an initial training stage, rewards and absence of rewards would signal to subjects (humans, but other primates would be preferable) allowed versus not allowed moves. Then, once they don’t make errors (above a low threshold) in moving the disks, problems of varying degrees of difficulty are presented. Time to evaluate and execute actions that lead from any starting configuration to any goal configuration would be recorded, and then averaged to show the learning curve as in figure 17b.
Hypothesis 1: planning and action chunks

The comparison with a non chunking mechanism would be possible by disposing of evaluative times from subjects with PFC or Basal Ganglia impairments, consistently with our general framework.
Chapter 7

Hypothesis 2: planning and internal model formation

Let us suppose, for instance, that the mnemonic image wished for by a child is the image of the mother’s breast and a front view of its nipple, and that the first perception is a side view of the same object, without the nipple. In the child’s memory there is an experience, made by chance in the course of sucking, that with a particular head-movement the front image turns into the side image. The side image which is now seen leads to the image of the head movement; an experiment shows that its counterpart must be carried out, and the perception of the front view is achieved.

*Sigmund Freud, Project for a scientific psychology (1895)*

In this chapter, first of all, I discuss how the brain has been described as a device processing information about rewards and punishments. Punishments can be considered as absence of reward. This motivates the use of Reinforcement Learning (RL) (Sutton and Barto 1998, see chapter 6 of this thesis) as a biologically plausible mathematical framework to model basic computations in the brain.

Second of all, I review some experiments and models in Instrumental Conditioning, based on rewards, that theories of PFC activity should take into account, in particular animal learning studies on Skinnerian stimulus-response habit mechanisms, producing the so-called superstitious behaviour (Skinner 1948). A simulation using RL to model this kind of behaviour will be illustrated.
Hypothesis 2: planning and internal model formation

Third of all, I show how superstitious behaviour cannot fully explain goal directed behaviour as demonstrated experimentally in mammals. One fundamental missing aspect is the capacity in mammals to learn the contingency (causal relationship) between stimuli and actions, or, in other words, to know and evaluate the consequences (states of the environment) of their actions before they actually execute them, i.e. internal models. Stimulus-response habits can instead only detect temporal contiguity. The capacity to "understand" causal relationships presupposes the use of an internal model of the world that predicts the evolution of the environment through time, depending on the executed actions. To describe directives of mammalian behaviour, rewards and punishments are not enough, internal models are needed too. This “understanding” should not be necessarily viewed as a conscious process. On the contrary, this view of Instrumental Conditioning is coincident with what I have called implicit planning. Lesion studies in the dorsolateral PFC of rats suggest that this is the part of the brain mainly involved in contingency learning. I present also the general lines for a simulation accounting for these lesion studies within my general framework for planning. These ideas clarify what is the specific role on the dorsolateral PFC in planning in respect to basal ganglia, namely to construct novel internal models based on contingency.

Fourth of all, I show other experimental evidence that the PFC is involved in learning internal models and storing them for some time until, through striatum, they are progressively stored in posterior cortex. In a final simulation, I model the temporal aspect of internal model prediction, as recorded in Watanabe (1996), in particular relative to temporal expectation of the final outcomes of actions.

7.1 Dopamine and reward based neural computations

A subjective feeling of pleasure or pain is how we individually experience respectively rewards and punishments. But this subjective feeling cannot be easily investigated scientifically, at least in animals. Therefore rewards and punishment have been studied in Classical (or Pavlovian) and Instrumental (or Operant) Conditioning as stimuli that act as positive reinforcers, if they increase the frequency of the
behaviour leading towards them, or negative reinforcers, if they decrease the frequency of the behaviour leading towards them.

In Classical Conditioning the increasing or decreasing of a certain behaviour in relationship to a reinforcer does not depend on actions performed by the agents. The behaviour is more like a re-action, e.g., salivation; whereas in Instrumental Conditioning, only specific actions can lead to specific reinforcers, e.g., press a lever. The capacity to process correctly and efficiently information about rewards and punishments can lead to the fundamental capacity to predict their occurrence, evaluate their magnitude, and possibly generate a number of strategies to obtain or avoid them. In this thesis I focus on rewards only, and consider, as mentioned, punishments as an absence of rewards, or negative rewards, but it must be noted that the biological system that processes and evaluates information regarding negative reinforcers might involve areas or neurotransmitters other than those involved in the reward circuit (as for example suggested in Daw et al. 2001).

The basic tendency towards rewards, modulated by a constantly changing motivational value, measuring the amount of effort that an agent will employ as based on current needs, seems a basic requirement for every living system attempting to survive. Note that motivation is an important aspect of Instrumental Conditioning, as discussed in Balleine and Dickinson 1998, and modelled in Dayan and Dickinson 2001, and it is relevant to ventromedial PFC activity. The mechanism of reinforcing an association between a stimulus and an action, if this action has led to a reward when the stimulus was presented, seems a necessary characteristic of adaptation. This mechanism has been even elevated to the status of biological law by Thorndike at the beginning of last century, the so-called “law of effect” (see Dennett 1979 for a discussion).

Several techniques permit us to assess if the brain is a processor of information about rewards: brain lesions, pharmacology, electrical self-stimulation, cellular recordings, neuroimaging. Several areas seem to respond to the delivery of rewards, for example the striatum, the dorsolateral and orbital PFC cortex, anterior cingulate cortex, amygdala and hypotalamus (see Schultz 2000 for a detailed review), but the major objects of study have been dopaminergic cells in the substantia nigra and in the
ventral tegmental area, because the activity of these neurons has been found to be of particular relevance to the actual processing of reward information.

Schultz and his colleagues (see Schultz 1998 for a complete review about this experiments) study monkeys and their brain reactions to rewards delivery, mostly through Conditioning, both Classical and Instrumental. A special phasic activity of the dopaminergic neurons (about 80%) in ventral tegmental area and substantia nigra pars compacta has been recorded, as illustrated in Fig. 17.

In the Figure there are three plots, each of them consisting of a lower part, showing the spiking activity of a specific neuron in different sessions of the experiment (time
Hypothesis 2: planning and internal model formation

Hypothesis 2: planning and internal model formation

goes from left to right), and an upper part, which is the sum of the activities reported in the lower part. The time course of this phasic activation is tens of milliseconds. In the mammalian brain normally there are modifications of the dopamine levels that take place over seconds and minutes (e.g., Louilot 1986), so fast specific kind of recording is probably looking only at just one of many possible roles of dopamine.

In the typical experiment to assess this phasic activity, there is a first stage where the monkey receives an unexpected reward at a certain time (R in the figure). The first plot of the figure shows the dopaminergic neuron activity at this first stage of learning: if the reward arrives, and it was not predicted, there is a phasic burst of activity immediately afterwards. In the second stage of the experiment, the monkey is consistently trained to see a Conditioned Stimulus (CS) some fixed time (1.5 seconds in the specific experiment reported in Fig. 1) before the reward is actually delivered. As showed in the second plot, after the monkey learns this CS-R association, dopamine is emitted immediately after the occurrence of the stimulus, and not after the delivery of the reward. The third plot shows the neural activity when the reward, that was expected because of the second stage, is not actually delivered to the monkey: a depression appears.

The same experiment can be conducted in an Instrumental version, where the delivery or not of the reward depends on the correct execution by the monkey of specific simple actions. Later in this chapter I describe a simulation of this form of the experiment.

One interpretation of the role of dopamine in this kind of experiments is to suggest that dopamine provides to brain structures (those receiving dopaminergic connections) a prediction error for rewards, that is to say the difference between the expected reward and the actual reward. This is exactly what is called a $\delta$ signal in the RL literature, described in the previous chapter. This interpretation suggests the hypothesis that some form of TD (temporal difference, see previous chapter) learning could be taking place in the brain, and regulate the strength of connections among neurons.

Montague et al. 1997 have suggested that the dorsal striatum, which is a part of the basal ganglia, together with dopaminergic release from the substantia nigra pars
compacta and its synapses within the striatum, might be implementing an actor/critic system (explained in the previous chapter) for the selection and sequencing of actions. The activity of dopamine would give the δ signal necessary to regulate the weights that decide which action to perform.

7.2 A simulation of the dopaminergic role in the basal ganglia

What Montague et al. 1996 have implemented is just the critic part of the system; what I illustrate here is a simulation that is an Instrumental extension of this system, that also includes the actor, with the process of having to perform immediately after a trigger two different sequences of actions depending on which stimulus was presented. The basic structure of my model is given in Figure 18. This simulation focuses on the following modules of the general framework for planning (illustrated in Fig 10 in chapter 4): basal ganglia (not as a chunk labelling module, but only as a reward processor) and the afferent dopaminergic connections, parietal and temporal cortex module, motor and premotor cortex.

![Diagram](image)

Figure 18

Hypothesis 2: planning and internal model formation

The actual implementation is a simple connectionist network as in Montague et al. 1996. The units are binary, and they can represent single stimuli, single actions and rewards. Time is represented as in Complete Serial Compound, i.e. stimuli, rewards and actions are represented as unique units for each different time steps. Each of these units has different (sending and/or receiving) weights, that are progressively updated by the learning processes based on TD, in the form of the equations given in section 3 of Chapter 6.

In figure 18 it is shown that in the model there two allowed actions (1 and 2), three different stimuli (1, 2 and trigger). Weights from stimuli to actions are symbolized as \( v(#,\#,t) \), where the first number represents the sending stimulus unit, the second number represents to receiving action unit, and the \( t \) is the specific time step, that, as required by Complete Serial Compound, points to the fact that the weights are different for different time steps. Weights from stimuli to the critic part of the model are symbolized as \( w(#,t) \), where the first number represents the sending stimulus unit, and \( t \) is, as before, a different time step. The units \( P(t-1) \) and \( P(t) \) compute the difference in the expectation of the reward delivery at two consecutive time step, according to the given TD definitions. The unit that computes \( P(t) \) also takes the current value of the incoming real (not necessarily expected) reward, thus having all the values to calculate delta at time \( t \). An error signal \( \delta \) is emitted at each time step, that adjusts the weights according to the gradient descent rule given previously.

In the following figures we can see how the simulation progresses. These figures contain six plots. Each trial of the simulation is divided into 8 time steps. In the plots, the current time step is always on the left, and, as the experiment progresses, and a new time step advances, the previous time steps are shifted progressively to the right, up to a total of eight time steps.

The simulated experiment has two different kinds of trials which the system has to learn, and they are presented randomly alternated. They consist of:

- a Stimulus 1 (coloured blue in the figure) for one kind, or Stimulus 2 (coloured in green in the figure) for the other kind, both at time step 1;
Hypothesis 2: planning and internal model formation

- a trigger stimulus at time step 5 (represented in red);

- a sequence of two relevant actions to be performed at time steps 6 and 7 (the other actions are not taken into account);

- finally a reward at time step 7, if the correct sequence of actions was performed (in yellow).

At the beginning of the experiment, a random behaviour is encouraged, so to have an exploratory search of the best policy.

The top left plot shows stimuli and trigger presentations in each consecutive time step, whereas the top right panel shows the actions taken by the agent. The middle left plot shows the weights from the stimuli and trigger (at different time steps, as in the Complete Serial Compound representation) to the P(t-1) unit, whereas the middle right and bottom right show respectively the weights from the stimuli and trigger to the Action units. Finally, the bottom left panel shows the history, for each experiment, of the values assumed at different time steps by P(t-1), P(t), Reward and δ (respectively coloured in light blue, dark blue, yellow and red). At the bottom of this panel it is also showed the trial number, corresponding at what stage of the training the agent is.

In the first trial (in Fig. 19) a Stimulus 1 is presented at time step 1, and the agent performs Action 1, given the initial random weight.
As learning progresses, and rewards are cumulated when correct actions are performed after the trigger, the randomness in the action execution, which is defined inversely proportional to the number of rewards, is reduced, and so the weights, as they have been adjusted by the TD mechanism, constraints more and more the actual behaviour.

During early trials the error signal is positive after the delivery of the reward, which occurs when the system randomly performs the correct sequence of actions, as can be seen in Fig. 20.
When the learning of the correct sequence of actions takes place, progressively two things happen. First, the exploratory behaviour is diminished, and I simulated this by a sigmoid progressively decreasing function of the randomness in the action choice (in figure 21 the case for the noise relative to the presentation of stimulus 1).

The second thing that happens is that the error progressively shifts backwards in time up to the occurrence of the first predictive stimulus, and, according to which of the stimuli has been presented, different sequences of actions are performed. This can be seen respectively in Figure 22 and Figure 23 for both different stimuli.
Hypothesis 2: planning and internal model formation

Figure 21
A sigmoid progressively decreasing function of the randomness in action choice, depending on cumulated rewards (the higher the total rewards, the lower the exploration). Red squares represent top possible randomness in probability of choosing the action, green diamonds represent actual values.

Figure 22
Error progressively (after 100 trials in this case) shifts backwards in time up to the occurrence of the first predictive stimulus (blue stimulus).
Figure 23
Error progressively (after 101 trials in this case) shifts backwards in time up to the occurrence of the first predictive stimulus (green stimulus).

The entire shifting backwards of the error signal (or dopamine emission) for each of the two different initial stimuli can be seen in the plot in Figure 24 and 25 (next page). Note that, even though TD methods to simulate dopaminergic emission capture some key experimental findings, there is a fundamental aspect of these models that does not match well with the actual recordings, as those illustrated in fig 17. In the data, the shift of the dopaminergic signal backwards in time from immediately after unexpected reward delivery (in early trials) back to the first predictive stimulus (in later trials, after learning has occurred), takes place as a sudden shift, without dopamine delivery during the time interval between reward and stimulus occurrence. This is not the pattern of the error signal in the TD models present in the literature,
Hypothesis 2: planning and internal model formation

where instead the δ shifts backwards in time over repeated trials and gradually, through all intermediate steps in the reward and conditioned stimulus (discretized) time interval.

The way in which this difference is made compatible with the data (Montague et al. 1996, for example) is by inserting a threshold in the dopaminergic cells, such that below it dopamine is not released. Given the u-shaped form that the delta signal takes through time, by setting the threshold below the two peaks, dopamine would be delivered only at reward time, for early trials, or at predictive stimulus time, after learning has occurred. Thus all the intermediary activations would disappear. Such explanation is not entirely satisfactory, because, even though dopamine would not be emitted, still it would have to be coded somewhere, in order for the TD mechanism to work. Therefore, with this explanation, the basic problem is only hidden, but I has not disappeared. Completely different modelling approaches, not based on TD, but on more complex anatomical and physiological considerations (Brown et al. 1999), don’t suffer of this (widely spread in the computational neuroscience literature) mismatch with cellular recordings, still waiting for better explanations.

The biological signalling of errors in the prediction of rewards as expressed by dopamine is nonetheless a secure and promising experimental finding for theoretical investigations. Not only to try to model, as we have seen, the working of the basal ganglia, but, given that the other locus of the brain that is massively targeted by the ventral tegmental area is the PFC, it deserves to be investigated if a simple signal like the error in the prediction of reward can explain or at least start to explain some of the cognitive mechanism in which the PFC is thought to be involved.

Given the general interpretation of the role of phasic dopaminergic activity towards the basal ganglia as an error signal in the prediction of rewards or novel stimuli, what is the role of this signal in respect to PFC activity?

In the third chapter we have seen how in high level connectionist models as developed in Cohen et al. (1996), dopamine in PFC has been modelled as affecting the slope, or
Hypothesis 2: planning and internal model formation

Figure 24
The entire shifting backwards of the error signal (or dopamine emission) for one of the stimuli (blue).
gain, of the units in the Context pools, and making the logistic activation function smoother or sharper. In this way, in the presence of dopamine, excitatory and inhibitory influences are stronger; in other words, as it often said, its role can be described as increasing the signal to noise ratio. Dopamine makes active states in PFC, as driven by other brain afferents, more active and stable, whereas low active states are made less probable to shift to higher activity states. This role has also been modelled as gating the flow of activity in PFC (see Figure 26) (Braver et al. 2000). This last study is actually especially relevant to the model I am developing here. Their
work is a further extension of J. Cohen’s model summarized in chapter 4, with the further inclusion of reward based TD learning mechanisms.

The authors extend their previous model by adding a Reward Prediction module, acting as dopaminergic activity, that computes the difference between the actual reward and the expected reward, as in TD learning. This module potentiates the strength of the excitatory connections from stimuli to context, and also the local inhibition from the tonically active bias unit within the context module. This double gating effect of dopamine has the role, as learning takes place, of favouring the activity of Context units that are the best predictors of rewards. Active maintenance of context information is therefore obtained using a dopaminergic gating signal that is triggered by error in the prediction of rewards.

Plenty of work, both at the theoretical and experimental level, remains to be done before we can say that we fully understand how rewards are processed in the brain, and what is the role of dopamine, both in subcortical and cortical structures (see also section 7.5). But in the rest of this chapter I want to go one step further, and question
the idea that reinforcers are all that is needed to learn goal oriented behaviour. Even though reinforcers seem to be necessary to an agent in order to produce behaviour useful to its survival, can we say that they are also sufficient to explain the variety of goal oriented behaviour we observe in mammals?

7.3 Contiguity versus Contingency

Is the “law of effect” (Dennett 1979) enough to model goal directed behaviour in mammals? That is to say, is the strengthening of the association of a stimulus and an action, after a reward is presented, sufficient to explain Instrumental Conditioning of rats, for example?

Skinner, a behaviourist, did think, in his 1948 Superstition in the pigeon, that:

“... conditioning takes place presumably because of the temporal relation only, expressed in terms of the order and proximity of response and reinforcement”.

He studied birds, and indeed the law of effect seemed enough to describe the behaviour he observed. He put a number of pigeons in boxes that were capable of automatically delivering rewards (food) in limited amounts and at different random times. Skinner left these birds in the boxes for some hours, and when he came back he found that most of them repeatedly performed a single action (like moving the head), different for each of the pigeons.

He interpreted this bizarre behaviour as due to the fact that when the first random reward came to each of the birds, they were by chance performing a certain action, while exploring the box. This random synchronicity in the two events led to a reinforcement in the execution of that specific action, in order to obtain the reward again. If further reinforced, this could lead to a “superstitious” behaviour and obsessive repetition of the action. He wrote:

“The experiment might be said to demonstrate a sort of superstition. The bird behaves as if there were a causal relation between its behaviour and the presentation of food,
although such a relation is lacking. There are many analogies in human behaviour. Rituals for changing one’s luck at cards are good examples. A few accidental connections between a ritual and favourable consequences suffice to set up and maintain the behaviour in spite of many unreinforced instances.”

But is the law of effect enough to describe what really happens in animals that have a more complex nervous system than bird, as in even simple mammals? The problem is (Balleine and Dickinson 1998) that this law only looks at the contiguous pairing of actions and reward, under the context established by the stimulus, but it does not look into possible causal relationships between them. In order to have a fully “Instrumental” Conditioning, with the capacity of the animal to influence events in the environment, we would expect an agent to know if its actions are the cause of some of the external events. As Balleine and Dickinson well summarize in their paper (page 408), an agent behaving only on the basis of the law of effect:

“… does not know about the consequences of its behaviour and thus cannot evaluate different courses of action in terms of the relevance of the outcomes to its current needs and motivational states.”

As an aside, note how in English we say that we “understand” something and act accordingly when we have “knowledge” of something. This is opposed to a “superstitious” behaviour, as also in animal learning, where instead the acts have no causal explanation. Interestingly, “understanding” means literally “standing under”, whereas “superstition”, from the Latin, means “standing above”.

Experiments to measure the sensitivity of mammals to causal relationships between reinforcements and actions have been described for the first time in Hammond (1980), and improved after that. In one experiment, reported in Balleine and Dickinson (1998), the authors pretrained hungry rats to press a lever and pull a chain in separate sessions, with the first action rewarded with food pellets and the second with a starch solution. Each reinforcers was delivered with a probability of 0.05 following each second containing at least one of the two actions. The causal relationship between
action and reward was then manipulated by arranging for one of the foods (the pellets for half of the rats, and the starch solution for the other half) a delivery with probability 0.05 following each second without an action. In this way the action still kept its previously learned effectiveness, but the probability of getting a reward after each second was the same with or without the execution of the action. Therefore, for each animal, one action (the one whose contingency has been manipulated) is rendered causally ineffective, because also not performing it produces the reward with the same probability; the other action, instead, remain causally effective and must be executed to get the reward. Figure 27 shows how the rats are sensitive to the differential causal status of the two action, as soon as the manipulation starts.

![Graph showing the effect of contingency manipulation on behavioral actions](image)

It can be seen that the probability per minute of executing an action whose food delivery can be obtained also without executing it (labelled "Same") becomes less and less from trial to trial and from session to session. The other action (labelled "Diff") in the first session also declined a little, but with further training it stabilized at a level well above the "Same" action. In the right panel there are the results of a final test in which no reward was delivered anymore. All the rats performed the action that was
Hypothesis 2: planning and internal model formation

causally related to the reinforcement more than the other action. This experiment is interpreted as showing that Instrumental Conditioning in rats is sensitive not just to the temporal contiguity between actions and reinforcements, but also to their contingency, namely the causal relationship.

The authors experimented further, and they repeated the same experiment but, after the pretraining and before the manipulation, they lesioned specific parts of the brains of the rats. The result was that if they lesioned the prelimbic area of the PFC, an area that corresponds roughly to the dorsolateral PFC of monkeys and primates (Fuster 1989), then, after the contingency manipulation, the decline of the probability of execution affected both actions, and not just the manipulated one, as it was instead in the normal animals (See figure 28).

![Figure 28](image)

From Balleine and Dickinson 1998. Explained in the text.

Also in the final test, the time to extinguish the actions when no reward was given, was approximately the same. This incapacity to discriminate was not a general product of cortical disruption, because when the authors lesioned the insular cortex of intact animals, the discriminations was still obtained.

Several interpretations might be found to explain this failure to distinguish between the contingencies, but one point that seems to hold in any case is that dorsolateral PFC
is part of a circuit that is involved in the formation of these expectancies about the world, according to the actions that are taken. Since animals form these associations, we can think that they learn and use an “internal model” of the world, that computes predictions to simulate future sensory inputs and outcomes of actions.

Another important point that must be made is that if only reinforcements could explain the strengthening of certain actions in association with certain stimuli, how to explain sensory preconditioning as in Tolman (1932). This experiment consists in a first phase where a first neutral stimulus is paired with another neutral stimulus, CS1 + CS2; then there is a second phase where the first neutral stimulus is paired with and unconditioned stimulus (food), until a conditioned response is generated (salivation). Finally the second stimulus is presented, and the result is that it immediately generates the Conditioned Response. Latent learning has also been explored in unreinforced navigation (Blodget 1929). Mostly rats show that, if preliminary left to explore a maze, and then subsequently rewarded by food put in a certain location, when put in a new starting location, immediately they are capable of going straight to the reward location. Tolman, and also Blodget, hypothesized that a large number of expectancies can be combined in a cognitive map independent of rewards. A cognitive map is a spatial way to instantiate an internal model, and we could say that it is an internal model of how the environment changes when no action is taken by the agent.

An accurate historical study should look into why early behaviourist studies did have these ideas on internal model formation in Instrumental Conditioning (Tolman, and his cognitive maps, but also in the work of C. K. Hull, e.g. in Hull 1984)), whereas later studies (as in B. F. Skinner) didn’t. A possible explanation for this is that experiments where you can control the signal and the reinforcer are easier to set up than if you leave the choice of the instrumental association to the subjects. And that is why historically there has been a shift to Pavlovian Conditioning instead of really studying Instrumental Conditioning in its full complexity and implications.

This full version of Instrumental Conditioning, with a reward based mechanism plus the construction of internal models, is a clear conceptualization of the idea of a basic unconscious planning process that I have discussed in chapter 5. Indeed it coincides with the idea of a purposive biological algorithm, that evaluates actions in the view of
achieving a goal, and that is based on the capacity to simulate the evolution of the environment depending on the actions that are taken (including no action). This reward-plus-knowledge mechanism can probably be found in all mammals, but in humans it is further enhanced by a complex capacity to communicate and share this knowledge with other members of their species, thus opening the mechanism to consciousness (again, see chapter 5 for a more detailed discussion).

In conclusion of this section, let me introduce the general lines for another simulation regarding the key experiment reported in Balleine and Dickinson 1998, and described earlier. I have not actually implemented this simulation, but a general description and a discussion using the basic hypotheses of the general model will suffice.

In the original experiment it is tested the basic idea of internal models, not just for temporal expectations, but more in general for the outcome prediction of different sets of actions. The actions that are available in the simulation are press_the_lever, pull_the_chain and no_action.

In the pretraining phase, if the animal presses the lever in an interval of a second, it will get food with a 0.05 percent probability; if it pulls the chain, it will get a drink with a 0.05 percent probability; if it doesn’t act, it won’t get any food or drink. Therefore, the animal will have to form the following internal models, that go from any state of the experimental context to a state where there is food or drink with a certain probability:

<table>
<thead>
<tr>
<th>Press_the_lever</th>
<th>New food</th>
<th>New drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any state</td>
<td>0.05</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pull_the_chain</th>
<th>New food</th>
<th>New drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any state</td>
<td>0.0</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No_action</th>
<th>New food</th>
<th>New drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any state</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
During the training phase, instead, also the no_action will bring a reward, food or drink in different animals, therefore an updated internal model for one animal, for example, will be:

<table>
<thead>
<tr>
<th>No_action</th>
<th>New food</th>
<th>New drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any state</td>
<td>0.05</td>
<td>0.0</td>
</tr>
</tbody>
</table>

while the other two internal models will remain the same.

The point of the experiment was to show that animals with intact PFC would progressively, during the training phase, extinguish the action that leads to reward with the same probability of the no_actions, simply because it becomes pointless.

This point, within the expressed hypotheses on planning, involves the internal model formation in the dorsolateral PFC, and also the evaluation process of which of the three task related actions is better to perform. If the goal is new food, and of the three available actions the no_action leads to it with a probability equal or even higher than any other action, this is the one that will bring to an higher evaluation of the current state, and therefore this is the one that is possibly executed. In sum, our views on planning fully support this aspect of the experiment.

When we try to model the disruption aspect, we must hope to produce the same errors in performances that are found in lesioned rats. But a fundamental problem that must be taken into account for this (and which is not reported in the paper) is that we should know how much of the learning process, in the pretraining phase, has become automatic, i.e. has been shifted to posterior striatum and Cortex. The time course for the automatization process varies from individual to individual, from species to species, and depending on how much and for how long the animals are pretrained. Therefore findings after lesions of the homologous of the dorsolateral PFC in small numbers of animals might vary considerably. If the automatization process has taken place at the time that the lesion is made, I would expect that the lesion induces the animal to persevere in the execution of the two actions even though one of them has become pointless, because a new task related internal model cannot be learned, and so only old automatized behaviour can be executed. If instead the shift to posterior
striatum and cortex has not taken place yet at the time of the lesion, I would expect that the internal model representation for the three actions in the pretraining phase is disrupted and extinguished, and simply the animals takes the reward when it sees it. The latter is the actual finding in Balleine and Dickinson 1998 reported experiment, therefore I would think that in those animals the shift to posterior striatum and cortex had not taken place at the time that they were lesioned.

7.4 Internal models in the brain

Balleine and Dickinson 1998 experiment does suggest that internal models find in dorsolateral PFC a key region for their formation. But what else can we find in the literature that relates internal model formation to brain structures. Besides the idea of cognitive maps developed and experimented in rats by Tolman in the '30s, which is restricted to navigational issues, it's in the '40s that the idea of internal model appears in full. The Cambridge psychologist K. J. W. Craik wrote (1943, p. 61)

If the organism carries a “small scale model” of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best for them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to reach a much fuller, safer, and more competent manner to the emergencies which face it.”

A simple but powerful idea of an output predictor has been used to describe the working of the nervous system for the purpose of motor control, and in particular of the Cerebellum as a repository of Inverse and Forward models. (for example in Wolpert et al. 1998, Kawato 1999, or more mathematically sophisticated versions using Kalman filters for sensorimotor integration in Wolpert, Ghahramani, Jordan 1995).

The idea of internal model has been formalized in several domains, most of all in Control Theory, and in particular in internal model Theory. Consider the system in Figure 29.
Suppose that $G_p^{-}(s)$ is a model of the process $G_p(s)$. Suppose further that the controller is a perfect inverse model of the process (as it would happen in the case of perfect control), i.e. a system that if given the goal as an input, as an output produces the input that is needed to the process to have the goal as an output. This can be written as $G_c(s) = G_p^{-}(s)^{-1}$. And suppose that the process follows entirely the model (perfect knowledge), namely: $G_p(s) = G_p^{-}(s)$. In this case the output will always be equal to the goal, and the controller does not require any feedback.

But these perfect conditions are not generally given, and besides there might be a variable disturbance $D(s)$ from the environments acting on the output of the system (see Fig. 30). This brings to the need of an inverse model (described before), but also of a forward model, that is an internal simulation of the external process. Given a controlling action $U(s)$, the comparison between the real output of the process and the simulated output, brings to the signal $D^{\wedge}(s)$, which is recursively fed back as a negative signal to the input. These basic ideas are used in Kawato (1999) models.

Figure 29
A typical schematic representation of a control process

Figure 30
A control process that uses an internal model to predict the outcome of its actions
Internal models have also been used to define formal adaptive systems that make use of a distal teacher (Jordan and Rumelhart 1992), as a mechanism that transforms distal sensory information about the consequences of actions into information for making internal representational modifications.

N. Schmajuk and Thieme (1992) have modelled latent learning with a neural network composed of a purposive action system controlled by a motivational system, and a cognitive system that constructs the cognitive map of the maze (i.e., the internal model). Their network is capable of fast time and real time predictions. The fast information in the cognitive maps is run "internally" to decide quickly what to do before the real execution. Their simulations match experiments in animal learning. While they think that the hippocampus plays the role of storing the cognitive maps, the PFC might participate in the purposive action system and in the generation of fast time predictions and the evaluation of the consequences associated with the choices.

More specifically with regard to PFC, the first to formulate the thesis, although in an implicit form, that the role frontal lobes might be to construct internal representations of the environment in order to predict its changes is in Miller, Galanter and Pribram (1960), as reported in Chapter 5 of this thesis. More explicit was the neuroanatomist W. J. H. Nauta (1971). He wrote:

"It is tempting to speculate that the reciprocal fronto-limbic relationship could be centrally involved in the phenomenon of behavioural anticipation, and elucidate the "loss of foresight" that has long been recognized as one of the most disabling consequences of massive frontal lobe lesions. The normal individual decides upon a particular course of action by a thought process in which a larger or smaller number of strategic alternatives are compared."

Nauta also discusses the role that affective components have relatively to the decision making process, which makes his speculations very similar to the more recent approach of A. Damasio (1994). Nauta didn't formalize his speculation, and he also seems to assume that the anticipation process (essential part of a planning process) is conscious.
The most detailed computational study of internal models in the brain can be found in Suri (2001). His model focussed on the so called anticipatory neural activity. Many brain areas have been shown to exhibit anticipatory neural activity, namely activity that, after learning, predicts the occurrence of specific events or rewards one or several seconds in advance (Hikosaka 1989, Apicella et al. 1992, Kermadi et al. 1995, Watanabe 1996, Hollerman and Schultz 1998). This activity is specific of the predicted event, and is not sensitive to specific characteristics of the predictive ones. R. E. Suri pairs this anticipatory neural activity to the processing of internal models. In his simulations he adopts TD learning and an approach to internal models that has also been used in Sutton and Pinette (1985). In his model, predictions \( p(t) \) and prediction errors \( e(t) \) are computed not only for reinforcements (as in the simulation explored at the beginning of the present chapter), but for all the presented events \( u(t) \) (i.e., stimuli and rewards). This \( u(t) \) is a vector that reports the presence or absence of events, which may last for some time. The basic idea of the internal model approach, which is the same I use in my simulations, is that the prediction errors are used to gradually update the weight matrix \( V \), that represents the “long term memory” of the model for any temporal representation \( x(t) \) of \( u(t) \), which comes from a special temporal transformation, consisting in a series of phasic signals covering trial duration. The first component of \( x(t) \) peaks with amplitude 1, the second with amplitude \( \delta \), the third with amplitude \( \delta^2 \), and so on. A multiplication of \( x(t) \) with \( V \) gives the state of the environment at the next time step \( x(t+1) \) discounted by a factor \( \gamma \). By iteratively computing this multiplication, as in Synchronous Value Iteration seen in chapter 6, the internal model can give a simulation of the environment over any time length. A comparison between what is predicted and what actually happens gives an error signal that updates the weights of \( V \) (an alternative approach to learn internal models is given in Dayan (1993) and the successor representation).

### 7.5 Internal model learning

As we have seen, there is plenty of experimental evidence that phasic dopaminergic activity is signalling an error in the prediction of rewards, and there is also plenty of
Hypothesis 2: planning and internal model formation

evidence that tonic neural activity in the striatum, as well as the PFC, is anticipatory neural activity in respect to rewards. But when it comes to error signals in the prediction of neutral stimuli (that might or not be subsequently useful to get rewards) or to anticipatory neural activity in respect to neutral stimuli (which again might or not be later used in reward acquisitions), both leading to the correct formation of internal models, the experimental research at the neural level is extremely rare. Balleine and Dickinson 1998 experiments do suggest that, at least in rats, the dorsolateral PFC might have a key role in the formation and storing of these internal models, but no clue is given on what is signalling to this brain area the errors in the predictions that allow their learning and updating. Is the error signal given by dopamine from ventral tegmental area? But isn’t dopamine related to rewards and not neutral stimuli? Can we hypothesize a role for dopamine in signalling an error in the prediction of every goal related event, both rewards and stimuli, as for example when associating the two neutral stimuli in sensory preconditioning? Or are these errors signalled by a completely different system?

The only alternative, and indeed interesting, proposal of which I am aware is given by Rao and Sejnowski (2000). In this work it is shown that biophysically realistic networks of recurrently connected cortical neurons can learn, through asymmetric Hebbian learning, internal models of the visual environment. The authors interpret feedforward connections from posterior to anterior cortical areas as carrying the error in the predictions of the visual stimuli, whereas the feedback connection from anterior to posterior cortical areas as carrying the prediction signals. Both views can be integrated, and we can hypothesize that while cortical-subcortical connections encode reward related information, and dopamine is the main carrier of an error signal regarding the predictions in reward occurrence, cortical-cortical recurrent connections instead encode stimuli-stimuli associations and internal models of the environment. In the model of planning that I adopt, I am not using Hebbian learning and the cortical-cortical recurrent connections. The expectations (temporal or object specific) of stimuli occurrence in my simulations will be indifferent to if the stimuli are rewards or neutral. This is equivalent to use the TD error signal not only for rewards, as in common models of dopamine activity, but also for any sensory
Hypothesis 2: planning and internal model formation

...event (as also in the simulations in Suri 2001). This approach is a first approximation, which makes the simulations easy to understand.

In E. Miller's laboratory it has been extensively studied how the PFC integrates and sustains information coming from ventral and dorsal pathways, respectively from inferotemporal and posterior parietal cortex, on what and where a goal related object is. Dorsolateral PFC receives more connections from the where pathway, whereas ventromedial PFC receives more connections from the what pathway, but strong connections within PFC suggest that the two kinds of information are put together. In a typical experiment, a monkey has to perform a delayed response task composed of two parts. In the first part the monkey, that is fixating the center of a screen, first sees a picture of an object; after a delay the same object and a new one are presented on the screen in different positions. Then there is another delay, after which the monkey has to make a saccadic eye movement to the remembered location of the matching object. They recorded neurons in PFC of the monkey, and they found that some neurons showed object tuned activity (what), some other location tuned (where) activity, and over half of the recorded neurons showed both what and where tuning throughout the delay periods. Similar results can be found in Goldman-Rakic's studies or J. Fuster's.

Under the light of computational requirements for planning and execution discussed previously, the interpretation which seems more correct is that the PFC of the monkey, when trained to perform the task through a schedule based on rewards, is constructing and sustaining an internal model of the task, relative to the specific action (i.e., the saccadic eye movement) that is planned. During the two delays periods, different neurons take care of keeping active the representation for what the object relevant to the action is (the object to match, which is actually stored probably in temporal cortex, but whose active maintenance takes place in the PFC), and where will be the location in the environment (again, the actual representation of the location is actually in the parietal lobes, but its active and task relevant maintenance takes place in the PFC) after the correct eye saccade is done.

The what information in this experiment is about an object identity that must be stored for some time until a certain action can be executed. Although it is about a past event,
it is in a certain sense future oriented information, because there is expectation that a certain object will appear in the visual field, and if it does, an action can be appropriately performed.

In Watanabe (1996) it has been showed more clearly how this what information should be regarded as the identity of the object that will be obtained in the future, and therefore it is an expectation. Watanabe recorded in monkeys’ brains while they were performing a delayed response task. The monkeys (Macaca fuscata) sited on a monkey chair, and were in front of a screen with a left and a right rectangular window, a left and a right button, and a central lever. The animals were trained to first press the lever for 6-8 seconds, then the left or right button was illuminated as a cue for 1 second. There was then a delay of 5 seconds, followed by a go signal (both buttons illuminated), and then the animals responded to the cued side in order to get a reward located in the window on the same side. The same kind of reward (a piece of raisin, sweet potato, cabbage or apple) was used for about 50 trials, but then it changed. Some of the neurons in dorsolateral PFC displayed spatial specificity (where information again), namely they showed activity related to which side (left or right) of the visual fields the cue was presented (Fig. 31).

Figure 31
Dorsolateral PFC cell displaying reward + spatial specificity, because it higher probability to fire when expecting a potato if showed a left light, but not if showed a right light (Watanabe 1996).
Some other neurons, as the one reported in Figure 32, didn't show this spatial specificity, but they were related to the specific kind of food that the animal was expecting.

![Apple](image)

**Figure 32**
Dorsolateral PFC cell displaying reward specificity only, insensitive to if the light was presented on the left or on the right (Watanabe 1996).

The *when* aspect appears also to be consistent with other findings in Watanabe's experiments. Some dorsolateral PFC neurons (as the one reported in Fig. 32) showed an anticipatory activity, that started at a low level at the beginning of the task (between the first and second vertical line), it grew maximally at the expected time of specific food occurrence, which of course coincides with reward occurrence (after the third vertical line), and then it decayed rapidly.

Watanabe interpreted the activity reported in the picture as reward specific, and without spatial specificity. I instead interpret this activity (in the light of ideas on TA and internal models) as the neural codification of a multi-time model as defined in a previous section. This neuron has no spatial specificity, but has instead objectual and temporal specificity, that is it codes when is highest the probability of a certain event (namely, the delivery of a specific object) to occur. This event also concludes a trial, and therefore it also concludes an option, in the RL sense. I show a simulation in this regard in the next chapter.
In Watanabe's experiment the stimulus (food) coincides with the reward, and therefore he underlines that these neurons are coding reward expectations, but my claim is that dorsolateral PFC can learn temporal expectations also of not directly rewarding stimuli, as long as they are task related, i.e. they are related indirectly to a reward. New experiments should test this hypothesis further.

The previous hypothesis is about what the representations in PFC are, i.e. Internal models, a second hypothesis is about the formation of these representations. It is reasonable to expect that several biological mechanism are involved in this learning process, but the present hypothesis focuses on the role of one specific neurotransmitter, which several findings indicate as a candidate indicator of an error in the prediction that the internal models make.

Recent biophysically realistic models of the role of dopamine in PFC (Durstewitz 2000 for a review) have described its effect as of enhancing afferent excitations into this area, and at the same time of lowering its spontaneous and endogenous activity. At the cognitive level, this mechanism seems to suggest a role of dopamine in sustaining and in actively maintaining patterns of activity in PFC, and at the same time in inhibiting spurious and distracting transitions to other patterns of activity.

In previous connectionist models (Braver 1999, 1999b, 2000) it has been suggested that phasic dopaminergic activity works as a gating signal, establishing when afferent inputs into the PFC should be maintained and kept in storage, or updated and changed, as described in a previous chapter.

One key issue when comparing biophysical and connectionist models is that while the dopaminergic role seem to be similar, in the former they are the tonic changes that are modelled, in the latter they are the phasic ones, acting at a much shorter time scale.

One attempt to reconcile the different views on which of the dopaminergic changes has an effect on PFC is proposed in Cohen et al. 2002. The idea is that while long term tonic changes might be mediated by D1 receptors, short term phasic changes might instead be mediated by D2 receptors. Besides, the two kinds of receptors, and consequently their effects, are proposed to be antagonistic. These two ideas (sustained by several experimental data not reported here, but summarized in the mentioned
Hypothesis 2: planning and internal model formation

paper) provide an integration of the role of tonic changes, long lasting and based on D1, versus phasic changes, acting on a faster time scale and enabling an updating role. The tonic dopaminergic effects, that instead increase the stability of the maintained representations, are not explicitly modelled here.

Another important aspect I have not addressed is that it is known that ventral tegmental area targets the PFC in an homogeneous way (Schultz 1998). Given this, recently some researchers (Frank et al. 2001) have preferred to identify the source of specific updating of PFC subregions in the connections from the basal ganglia and not in the connections from the Ventral Tegmental Area. I suggest that the integration of the model in with the chunking mechanism proposed here might be an interesting direction of research.

7.6 Simulation of temporal aspects of Watanabe 1996 experiment

In this section I describe a simulation of aspects of Watanabe 1996 experiment which relevant to my general framework for planning.

I tested the general planning framework specifically to see if the construction of multi-time internal models could match cellular recordings in dorsolateral PFC. I simulated Watanabe’s task to look into the internal models construction, but I didn’t implement the chunking process, also because it was not an issue in the original experiment. This simulation on the internal models can match part of the general behaviour of actual recordings in several neurons of the dorsolateral PFC. In terms of anatomical parts, in this specific simulation I am modelling the following part of the general framework (see Fig. 10 in chapter 4): dorsolateral PFC, and its afferent dopaminergic connections, reward evaluation in basal ganglia, and motor and premotor cortex module, posterior cortex.

This simulations mimics the architectural structure of the actor critic simulation seen in section 7.2, but the error signal are not just specific for rewards, but also for task relevant stimuli. Stimuli are therefore considered not for their rewarding value, but as salient perceptual outcome of actions.
Hypothesis 2: planning and internal model formation

The units of this simulation are binary, and they can represent single stimuli, single actions and rewards. Time is represented as in Complete Serial Compound, i.e. each of these units has different (sending and/or receiving) weights depending on the time step, and they are progressively updated by the learning processes based on TD, in the form of the equations given in section 3 of Chapter 6.

In the model (see Fig. 33a) there two allowed actions (left and right), and two different stimuli (left light and right light). The units P(t-1) and P(t) compute the difference in the expectation of the reward delivery at two consecutive time step, according to the given TD definitions, but also the expectation of other stimuli, in particular, of food. The unit that computes P(t) also takes the current value of the incoming real (not expected) reward or stimulus, thus having all the necessary values to calculate the errors in the prediction of their occurrence. The activity of the Dorsolateral PFC units in the model is therefore mainly influenced by the delta signal coming from the unit P(t), which is directed by stimulus occurrence and not just by reward, as in traditional TD learning. Dorsolateral PFC units, therefore, learn to predict the expected time of (rewarding) object occurrence, and to link this occurrence to the specific stimulus that is presented (left or right light), and to the specific action (left or right press) that is performed.

The specific formula used to calculate the error in the prediction of the occurrence of events (stimuli and rewards) is an extension of the one used for TD:

$$\delta_u(t) = u(t) + \gamma V_u(t+1) - V_u(t)$$

where $u$ is the specific event (three types of food, in this case), $\gamma$ is a discount factor, $V_u$ is the prediction relative to the specific event $u$, and $t$ is time. The decision function that establishes how exploratory should the behaviour be is sigmoidal (as in a previous simulation in this chapter, Figure 21), and dependent on the number of accumulated rewards, i.e. the higher the number of rewards, the lower is the probability that the system will randomly shift from one action to the other.

The key idea used here is that phasic, D2 based, dopaminergic effects from ventral tegmental area act in the PFC as an error signal in the prediction of every task related
Hypothesis 2: planning and internal model formation

stimulus, thus allowing the construction of internal models. This idea has also been used when making computational models that take into account latent learning (see chapter 6), as in Sutton and Pinette (1985), and Suri (2001). While in other model dopamine is modelled as signalling only errors in the prediction of rewards, here it is modelled as signalling an error in the prediction of any event, that is to say stimuli, reinforcers and responses.

In figure 33b we see the activity of the PFC unit, encoding the prediction relative to the probability of terminating the press left actions with the food (apple). What is showed is actually the learning process of this prediction, in the form of growing activations, trial after trial, up to a total of 50 trials. The higher curves show the prediction in later trials, and therefore are those that must be compared with Watanabe's graphs.
As in the biological recordings, in this unit we see that option termination, which is coincident with stimulus (food) or reward expectation, grows staidly until the time of delivery of the reward. This specific class of neurons in dorsolateral PFC is coding the temporal expectation of specific and task related stimulus to occur. In Watanabe's experiment the stimulus (food) coincides with the reward, and therefore he underlines that these neurons are coding reward expectations, but our claim is that dorsolateral PFC can learn temporal expectations also of not directly rewarding stimuli, as long as they are task related, i.e. they are related indirectly to a reward.

The specific $\gamma$ used for the plot in figure 33b is 0.9, and the actual values taken from the decision function are showed, trial by trial, in Figure 33c.

![Figure 33b](image)

**Figure 33b**
Discounted exponential object expectation. In the actual simulation I have divided the task into 11 discrete time steps (see Figure 33b). The blue vertical line stands for cue onset (left light, in this case); the red vertical line stands for the go signal; the yellow vertical line stands for reward delivery which coincides with food presentation. The reward is given to the agent only if the correct actions (press left after left cue and go, or press right after right cue and go) are performed.
The effect of $\gamma$ is to produce an exponentially discounted value function, as those represented in Figure 33b. The lower is $\gamma$, the higher will be the discount, and therefore also the curvature of the value function. In Figure 33d I have plotted the predictions in a simulation as in Figure 33b, but here each subplot has different values of $\gamma$, ranging from 0.1 to 1. It can indeed be observed how higher $\gamma$ have the effect of discounting less, and of making the predictions more linear, or even to invert to curvature of the value function.
Figure 33d

Same kind of plot as in Figure 33b, but with 10 different values of $\gamma$. With lower values, the predictions of the occurrence of the stimulus are more discounted.
Hypothesis 2: planning and internal model formation

In the next two figures I have plotted again stimulus expectation and decision function in the same situation as before, but for 150 trials, instead of just 50, and \( \gamma = 0.7 \). It can be seen how the exponentially discounted shape is not affected by the number of trials, as long as the value of \( \gamma \) is appropriately chosen (i.e., discounting enough).

![Graph showing stimulus expectation and decision function](image)

Figure 33e
Learning of the value function predicting the occurrence of an apple if a press left action is performed during a sequence on 150 trials. The value of \( \gamma \) is 0.7.

In conclusion, in this chapter and with these simulations I have focussed on the following points:
Hypothesis 2: planning and internal model formation

- Part of the brain activity in animals is devoted to computing rewards and predictions of rewards in order to improve behavioural performance.

- At least in mammals, goal directed behaviour cannot be uniquely explained using reward based information, but also stimulus-stimulus and action-reward associations. In order words, a fully Instrumental Conditioning must hypothesize that the animals develop and use internal models of the environment. This more proper Instrumental Conditioning coincides with implicit planning.

![Figure 33f](image)

Value taken by the decision function for the same experiment as in Figure 33e. Red squares represent maximal possible value for the noise in action choice in each trials, blue diamonds are actual values of noise produced.

- Several brain areas might be encoding these internal models. Dorsolateral PFC seem to be somehow involved in the formation of such internal models.
Hypothesis 2: planning and internal model formation (Balleine and Dickinson 1998). Posterior regions of the cortex and cerebellum might be the place where these models are stored and updated.

- Precise computational models, both at the neural and connectionist level, can help to understand how the brain might actually implement the reward information processing and the internal model formation.

- RL seems to be a natural formalization paradigm, and its definition of internal model possibly gives a precise description of what working memory representations might actually be, i.e. what is actually stored in short term memory are not spatial or sensorial representation *per se*, unrelated to time, but instead predictions of *when*, *where* and *how* the environment might change according to the actions that are executed.
Chapter 8

Conclusion: discussion and future directions

This chapter constitutes the conclusion of the present work, even though it is clearly not the end of the subject. PFC and planning as its role in the organization of behaviour are a research topic that cannot be exhausted in a dissertation. Nonetheless there are a number of fundamental points underlined in this thesis that I hope to have shown to be an advancement in the understanding of PFC and its role. These points are:

- Planning is the best candidate for a unifying interpretation of the role of PFC in behaviour.

- Planning can be defined as the goal oriented capacity of a system to anticipate and evaluate possible evolutions of the environment depending on the actions that are executed. Note how *anticipation* refers to the process of expecting that something will happen, for example the appearance of a certain object, and it is in a certain sense a passive process, because it does not imply necessarily the preparation for action. The process of *planning* is instead an active process of generation of possible future scenarios, by means of anticipations, and of evaluation and selection of which actions is preferable to perform in order to reach a goal. *Internal models* are anticipations about the environment, linked to the execution of specific actions, i.e. they code the perceptual consequences of actions. Putting
together internal models, and evaluating how the final anticipated outcome is close to the actual final goal, constitutes a full planning process. A plan can then be possibly executed or not.

- (Hypothesis 0) Planning should not be viewed, as commonly happens in cognitive neuroscience, only as a conscious activity, explicit and communicable, but primarily as an unconscious activity, implicit and not communicable. Implicit planning is the work of an information processing device, such as the brain, that is trying to anticipate and evaluate that best course of actions to take in a physical world in order to achieve a goal, and improve the balances of the body. Implicit planning is not the basically linguistic process of reasoning about what's best to do in a social environment.

This hypothesis is clearly testable. It is quite possible to design experiments that show if humans are able to learn how to perform goal oriented sequences of actions if and only if they are also capable of making a verbal report of their plan (what they are doing, how they are doing it, and which is their goal). As at length discussed and illustrated in chapter 5, these experiments, though rare, exist, and it is clear that people in several situations do plan unconsciously. Furthermore, if it can be shown that particular areas of the brain are disproportionately active when humans learn implicitly to plan as opposed to when learning explicitly, then those areas are good candidate as coding unconscious plans. Berns et al. 1997 experiment goes in this direction, and right dorsolateral PFC seems such a candidate. As also reported in chapter 5, humans who have lesions in PFC might have deficits in planning consciously, and not unconsciously (as in some forms of apraxia), thus also indicating a dissociation in the two modes of planning.

- The PFC achieves this unconscious planning capacity in concert with other brain areas, i.e. mainly the basal ganglia, with an action chunking role (what I call Hypothesis 1).

The specialized role of the dorsolateral striatum in the basal ganglia is testable, because it is possible to construct experiment that show how this area is more
active when some compressed recoding of actions takes place, and it is also possible to lesion locally this are to see in there is an impairment in the execution of previously learned chunks of actions. These experiments exists, at least in rats and monkeys, and they have been discussed in chapter 6.

- The dorsolateral PFC constructs on line internal models of the world through a learning mechanism based mainly on dopaminergic error signals about when, what and where goal related events are expected to happen in the environment depending on the actions that are taken (Hypothesis 2). Posterior cortex and the cerebellum have instead the role to store permanently well acquired internal models, previously learned on line in PFC (what I call Hypothesis 3). This division of labour makes the planning capacity faster and faster, by simply omitting the re-evaluation of well know information on the consequences of behavioural habits, thus, together with the basal ganglia, constructing hierarchies of actions.

Both hypotheses 2 and 3 are testable. Tasks requiring the formation of novel internal models, that is to say anticipations of what, when and where certain events will take place in the environment (such as in delayed response tasks) should activate disproportionately specifically the PFC. Tasks requiring instead the usage of well acquired internal models should activate preferentially posterior cortex. Given the wide range of this hypotheses, only pieces of these experiments can at the moment be found in the literature, and I have discussed them in chapter 4 and 7. Further tests can be designed to test these hypotheses, as the one described in section 8.1 to test specifically temporal expectations.

- Reinforcement Learning is a simple mathematical framework within which the described planning capacity working in the brain can be easily simulated. In particular, its use of concepts like rewards, internal models, temporal difference learning and temporal abstraction, make Reinforcement Learning and its more recent advancements perfectly in line with the main concepts to be used to understand the described brain planning dynamics.
8.1 A proposed experiment

The points just discussed constitute a general framework for understanding previous results on PFC activity and for designing further tests. As an example, consider the following possible new experiment to test the construction of internal models, in particular for temporal expectations.

A monkey is trained on a machinery that has two buttons: a red button, that opens a little door with a weak reward (bread) after 5 seconds and for just the time to pick up the reward; a green button, that opens a different little door with a favourite reward (banana) after 10 seconds, again for a 1 second. Subsequently (or previously, it doesn’t really matter) the same monkey is trained on a different machinery (a big box), such that if a yellow light appears, then immediately after a big door opens up for 7 seconds; if instead a blue light appears, then immediately after the same big door opens for 12 seconds.

After these two training phases, the monkey is tested on a trial where the first machinery is put behind the big door of the second: the monkey has to plan (if hungry enough) that if it sees a yellow light, when the big door opens up, it must press the red button, because it only has enough time to catch the less favourite reward. If instead a blue light appears, it can press the green button, because there is enough time to get the favourite reward.

This experiment is interesting because it involves planning, i.e. the monkey has to simulate and evaluate possible future courses of action (press either red or green button after it sees a blue or yellow light), and make a choice based on temporal expectations.

Our hypotheses predict that different neurons would encode different action related temporal expectations, so to have four classes of neurons, coding temporal information in the form of climbing activity:

1) Press red – 5 seconds – weak reward
2) Press green – 10 seconds – strong reward
3) Yellow light – 7 seconds – Big door closed
4) Blue light – 12 seconds – Big door closed

Cellular recordings in PFC *a la* Watanabe should give insight in this, and how the activity of the neurons would code such task. Presumably some neurons would code for events (i.e. door closing) and some other for specific reward expectancy (i.e. bread versus banana), both with sustained activity. Other neurons, according to my hypotheses, are expected to code temporal information in the form of climbing spiking activity culminating at the time of expected reward delivery.

A control experiment is to train different monkeys on a similar machinery, but by varying the temporal dimension by differently randomising, already in the training phase, all the opening time intervals of the three doors. The amount of “noise” injected into these time intervals should remain the same for each monkey, but vary from monkey to monkey. Animals trained with small amounts of noise (in the order of milliseconds), according to my hypothesis, are still expected to be able to predict the temporal aspects, possibly with slightly varying slopes in the climbing activity of the temporal related activity. But monkeys experiencing greater amounts of noise (in the order of several seconds) are still expected to code events and reward expectation by means of sustained activity, but they are not expected to code temporal information, because it would be so uncertain to be impossible to code, or useless for planning.

If the climbing activity of the kind described is observed in the (specifically dorsolateral) PFC in the no noise and in the low noise conditions, but not in the high noise condition, we can say that the hypothesis that PFC neurons code for temporal expectations in the service of planning is not falsified. Any different pattern would instead falsify it.

Also, there should be very different patterns of activity at the beginning of the testing phase and subsequent phases. In fact later on the monkeys, in the no noise or low noise conditions only, can just learn to associate yellow light with pressing red, and blue light with pressing green, without planning. In that phase no temporal related PFC activity could be observed.
The proposed novel experiment is just to give a sense of the several possible tests that can be generated within the given framework to describe PFC function.

8.2 Future directions

Our general model of neural planning, with an Internal Model construction module, an action chunking mechanism, and the capacity to shift action representations to other modules during skill acquisition, can match some actual neuronal recordings and, at the same time, achieve satisfactory planning efficiency, both in motor and intellectual domains. But several modifications to the general model can be imported and new directions need to be investigated. Here is a partial list:

- Different ways to learn Internal Models using TA can be configured: the exposed methods require an option to be executed until termination, others, more appealing from a psychological point of view, learn about an option from small fragments of experience consistent with that option, even if the option itself is not executed. These last methods are called *intra-option*, and they are a kind of learning that takes place *off-line*, i.e. the consequences are learned while potentially behaving according to another option. In these methods, options not executed and not relevant to the reward that the system is trying to maximize, can nonetheless be improved for a future use (if when searching for your keys, you also see that you left a specific book on the desk, you might need this information when you need that book).

- The previous problem is related also to the issue of how to deal with plan interruptions, that is to say how the nervous system is capable of interrupting the execution of a previously top ranked plan when circumstances change. This happens constantly, and it is reasonable to suppose that the brain, that has evolved to deal with an highly uncertain and mutable environment, has developed special mechanisms to work efficiently with interruptions, when necessary.
Another fundamental problem of the presented model is how to decide when the system should allow a sequence of actions to be chunked and become an option. This problem is actually a major subject of research in current RL studies (Precup 2000). In the simulations reported here the chunking mechanism comes into play indiscriminately, all the times a sequence of actions (or options) has been successful in reaching its goal at least once. Even though I have showed that this indiscriminate mechanism improves computationally the performance, it is obvious to think that more sophisticated procedures can be realized.

There are two basic ideas that appears relevant, both in computational and psychological terms. First, an agent should “start small” when learning about the environment (a similar discussion, but in a linguistic domain, can be found in Elman 1993), by setting simple goals, which have a short distance from the current state, so that these small options can be more easily part of bigger plans, as opposed to very long chunks that cannot be decomposed. Second, the chunking should come into play only for those sequences of actions that lead to goals after several successful repetitions, in order not to abuse of the computational capacities.

One way to achieve both aspects might be to consider that the brain might capable of computing how uncertainty (or predictability, which is its inverse) is present in a sequence of inputs that depend on which actions are executed. If the sequence goes below a certain threshold level of uncertainty (or, which is same, above a certain threshold of predictability), this might be a way to signal to the system that it is convenient to chunk this sequence of actions into a new unit. Uncertainty is precisely quantified in Information Theory (Cover and Thomas 1991) as entropy, i.e. the product of the probability that a certain events will occur time the logarithm of that probability. The higher this value, the higher is its unpredictability. If an event has a probability of happening of 0.5, it is highly uncertain (high entropy); if instead it approaches the 1 or 0 values, it becomes more and more predictable (low entropy). With this regard, a very important recent finding and potentially full of consequences (Fiorillo et al. 2003) regards the activity of dopaminergic neurons. Besides the well know activity signalling an
error in the prediction of reward occurrence, it has been a previously unobserved an independent sustained activity that covaries with (and thus indicates) uncertainty of reward occurrence. This might be a key signal to basal ganglia to start a chunking mechanism. In future work I will simulate these findings with the purpose of ameliorating the general planning framework.

- One important brain structure that I have left out of the picture is the thalamus. It is the gate of the inputs to the cortex as coming from the basal ganglia, and it is general more active when actions are actually executed, contrary to being just simulated. Future studies will have to include the thalamus in the general model of planning.

- What are the roles of all the neurotransmitters in the PFC? As in any other part of the cortex, several neurotransmitters are present in the PFC. In this thesis I have discussed only the possible roles of dopamine, but also acetilcoline is present, especially abundant in layers from III to VI, with projections coming from the basal forebrain cholinergic system. GABA is also found in the PFC, acting in wide-arbor and chandelier type interneurons, with an inhibitory role. A computational model of the PFC should propose hypotheses of the computational roles of these neurotransmitters, especially, as it is becoming increasingly evident in modelling of also other parts of the brain, their function in regulating learning.

- Are the roles of the left and the right PFC hemispheres different? The literature is not abundant on this point. I did report an experiment that does suggest (Berns et al. 1997) that the left part might be more related to goal oriented behaviour where language is involved and so subject can also verbally report on what they are reasoning, whereas the workings of the right part might be less conscious to the patient, and more spatial patterns might instead be processed.

- How can verbal information be passed to the model? What are the mechanism by which conscious planning is formed? Humans can be verbally instructed on what
to do and how to do it: this learning mechanism doesn't involve (at this stage) reinforcements, as instead in an operant conditioning framework, or progressive training, as instead in supervised framework (e.g. back-propagation). In fact, instructed learning can be immediate, as when rules of a game are given. Here symbols and rules have an immediate effect on the neural representation.

- Can long term plans (one day or more) be explained using this model? It seems that active maintenance of patterns of neural activity is not plausible: more likely they are stored somewhere and then activated at the appropriate time. The hippocampus, so fast in learning, might also be involved in this process. This structure might encode an association between the desired goal representation within the PFC and features of the circumstances under which the goal should be evoked. The detailed nature of such interactions, their relationship to the dopaminergic learning mechanisms just described, are still to be determined.

In conclusion, in this work we have seen that there is a coherent computational role we can assign to the PFC, which is consistent with experimental data, and which motivates a new understanding of the varieties of planning in cognition. We have outlined a general framework (including one novel experiment) to test this understanding, and have set the stage for a long term research program looking at the role of neural and implicit planning in cognition. We have tested two key aspects of this understanding through computational models, namely chunking and the construction of internal models for temporal expectations of salient events.
References


References


Craik, K., 1943: The nature of explanation. Cambridge University Press.


References


Jackendoff, R. S. 1990 Consciousness and the computational mind. The MIT Press.


References


Libet, B., Freeman, A., Sutherland, K., 1999: The volitional brain. Imprint Academic.


Robinson, Blair, Cipollotti 1998: Brain. 21, 77-89.


References


Uylings, H.B.M, van Eden, C.G., 1990, Qualitative and quantitative comparison of the prefrontal cortex in rat, and in primates, including humans. Progress and Brain Research. 85, 31-62.


