NEURAL MECHANISMS OF REINFORCEMENT LEARNING IN THE BASAL GANGLIA

Thesis submitted for the degree "Doctor of Philosophy"

By
Genela Morris

Submitted to the senate of the Hebrew University of Jerusalem

July 2006
This work was carried out under the supervision of

Prof. Hagai Bergman
Acknowledgements

My deepest gratitude is due to my advisor, Hagai Bergman, for encompassing far more than anyone could wish for in an advisor. His vast knowledge, analytical thought, experimental brilliance along with open-mindedness, absolute scientific integrity and undying curiosity are traits I could only hope to try to imitate. More importantly, Hagai's friendship was, and still is, invaluable. Eilon Vaadia has been like a second advisor to me. It is hard to imagine embarking on any stage of the experimental work without his advice. I was lucky to have Yifat Prut at close distance. Her help and support in both the technical and fundamental issues has saved me and my work countless times. I am also indebted to Moshe Abeles for his aid.

I would like to thank my many friends from the HBF workgroup. Yoram Ben Shaul, for introducing me to the work and the topics, luring me into staying and teaching me just about everything. Rony Paz, for endless discussions and arguments which have challenged me to improve. To both of you for being great friends. Dudi Arkadir, for being a wonderful partner to experiments. You made recordings immensely enjoyable. Alon Nevet, for setting an example of originality and courage of thought and for showing me the power of uncompromising analytical pursuit. It was your stubbornness that allowed me to record dopamine neurons. Gali Havazelet-Heimer, the only person I'll ever agree to share a 1 m² space with; Michal Rivlin, the smartest, most pleasant person to consult; Aeyal Raz, who generously shared his data; Josh Goldberg, Izhar Bar-Gad, Guy Saban, Shlomo Elias, Mikael Levy, Naama Parush, Mati Joshua, Boris Rosin for making the working environment simultaneously productive and fun.

During my long stay at the ICNC I was fortunate to be aided by two secretaries – Aliza Shadmi and Ruthy Suchi, who faithfully guided me through the bureaucracies of the PhD studies. Special thanks are also due to Esti Ben-Shoan from the physiology department, who is a truly invaluable asset.

I would like to thank the Horowitz Foundation for its generous support during three years of my studies and the Sally Berg foundation for its support.

My deepest gratitude is owed to my family. My parents, Luba and Yizhak; my brother Shmuel. I know you hate inventory lists of compliments, so I'll spare you – you know how I feel; Tony, for your love, and for mine; and my sons, Adam and Gideon for putting it all in perspective.
Table of contents

Acknowledgements ................................................................. 3
Abstract ................................................................................. 6
Introduction .............................................................................. 9
   I. Learning and reward – behavioural implications ............... 10
   II. Machine reinforcement learning ........................................ 14
   III. Physiological counterparts of reinforcement learning ...... 18
Methods .................................................................................. 25
   I. Subjects and training .......................................................... 25
   II. Surgery and recording ....................................................... 28
   III. Data collection and analysis ............................................ 31
Results .................................................................................... 37
   I. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons ......................... 37
   II. Implementation of temporal difference (TD) learning by dopamine neurons: implications for the computational algorithm ......................................................... 44
   III. Midbrain dopamine neurons encode decisions for future action ...................................................................... 50
   IV. Physiological studies of information processing in the normal and parkinsonian basal ganglia: pallidal activity in the Go/No-Go task and following MPTP treatment .......... 61
V. Anatomical funneling, sparse connectivity and redundacy reduction in the neural networks of the basal ganglia ................................................................. 70

Discussion ........................................................................................................ 79

I. Dopamine neurons reflect an accurate TD-error signal in the positive domain................................................................. 80

II. Monkeys use a probability-matching decision strategy ... 82

III. The decision strategy is shaped by dopamine neurons ... 83

IV. Dopamine neurons reflect future decisions ......................... 85

V. Striatal TANs report the occurrence of behaviourally significant events, Or: Dopamine-acetylcholine balance, revisited .................................................................................................. 88

VI. GP neurons may participate in action selection but not in action execution ................................................................................. 91

Reference List ................................................................................................. 93

(References are for all sections excluding Results I, III, IV and V which contain their own reference sections)
Abstract

Theory of human, animal and machine learning divides the concept of learning into two broad types characterized by the information that the learner is supplied by the environment. On one extreme, supervised learning deals with situations in which a teacher is provided to guide and correct performance. Implementation of such learning in neuronal circuits is obviously very costly and rarely feasible. On the other extreme are tasks that require no external feedback, since there is no ‘correct answer’. The learner just learns how to organize information in a manner which will ease its handling in further tasks. This limited set of learning problems is referred to as unsupervised learning. The psychological and computational field of reinforcement learning is situated at an intermediate step, in which the learning organism is left to explore all possible solutions (or actions) on its own, but limited feedback is provided regarding the suitability of the solution in the form of reward or punishment. Theorists in the field of artificial intelligence have studied this type of learning intensively. They have developed a powerful reinforcement learning algorithm known as temporal-difference (TD) learning, which overcomes the major difficulties of learning through unspecific reward. In this method of learning an estimate is made in each point in time regarding the expected reward that will follow. When external reward (or punishment) delivered, it is translated into an internal signal indicating whether each point in time is better or worse than predicted. This signal is called the TD error, and it serves to improve prediction and reinforce (or extinguish) particular behaviours.

Reinforcement learning, first studied by psychologists and later embraced by computational learning theorists, has lately been subject to physiological study in a search for a neural correlate of this type of learning in brains of humans and behaving animals. Early electrophysiological recordings have implicated two neuronal populations in signalling the TD error: dopamine (DA) carrying neurons in the SNe nucleus and acetylcholine (ACh) carrying neurons in the striatum. Interestingly, both populations target the same area: the striatum, which is the input nucleus of the basal ganglia, and specifically the connections between these inputs from the cerebral cortex and the projections out of the striatum to the subsequent basal ganglia stations. These connections are the cortico-striatal synapses. The basal ganglia, in turn, are
probably heavily involved in transforming external sensory information into the appropriate behaviour, or action selection. The aim of my research was to investigate the involvement of DA and ACh neurons in reinforcement learning, and in particular their role in action selection, and to incorporate these neuronal populations into a clearer and more comprehensive view of basal ganglia function in health and disease.

It was previously believed that both types of neurons provide the same message to the same target populations. This puzzling redundancy led me to the hypothesis that although the messages of the two are different and complementary for the task of learning, electrophysiological studies that have thus far been conducted were accompanied by behavioural tasks that could differentiate between them. I therefore designed a set of experiments to specifically address this question, and indeed these shed light on a dissociation between the two neuromodulatory systems: While DA neurons do provide an accurate TD signal (for positive rewards), ACh neurons provide no information regarding the rewarding value of external events. Rather, they provide a very accurate timing signal to the points in time where learning should take place most efficiently.

A chapter describing this result is followed by a chapter in which I describe the evolvement of the DA signal during subsequent behavioural epochs. This analysis serves to provide an estimate for the time-constants with which the cortico-striatal system can use the DA error signal to update the synaptic efficacies of its connections.

A key aspect of reinforcement learning and in fact of the behaviour of any adaptive creature (human or animal, alive or artificial) relies on its ability to perform decisions in the face of ambiguity. The third chapter of this dissertation deals with decision-making and the role of DA neurons in this complex task. In this chapter I show that the responses of DA neurons shape the long-term behavioural policy of behaving animals. I further show that the responses of these neurons that are exhibited as early as at the time of presentation of the dilemma, reflect the decision that is about to take place in the future, placing these neurons downstream of the decision junction.

A fourth chapter extends my analysis to further nuclei of the basal ganglia, and specifically to data recorded in the globus pallidus, which is the output nucleus of the
basal ganglia. This chapter shows how information regarding the chosen action is transmitted from the striatum to the 'operational' stations.

Finally, in the fifth chapter we will compare the various nuclei in which I recorded in terms of the information their respective neuronal populations hold regarding the actions that should be performed given some environmental situation, and the reward that should be expected following this action. This chapter illustrates an overall view of the basal-ganglia input and output nuclei, and proposes a comprehensive account of their function during learning of behaviour in view of environmental reinforcement.
Introduction

The field of computational neuroscience treads in a very delicate territory because of its inherent interdisciplinary nature. A dangerous pitfall awaits the zealous researcher in the form of the great temptation of unidirectional works. However, it is important to keep in mind that interdisciplinary science should be a tool to achieve the very same goals as all other classical disciplines have set to achieve. In my view, the field of computational neuroscience is only justifiable inasmuch as it can support a viable and recurrent dialog between its ‘computational’ and ‘neuronal’ counterparts. Thus, it can only thrive by studying structures in which the state of knowledge is such that the basic apparatus is in which three conditions are satisfied. First, the phenomenology of the function of the normal and pathological structure forms a clear enough image of the theoretical problem the structure is dedicated to solving. Second, the neuronal apparatus is sufficiently known in terms of physiology and connectivity, so as to be able to correctly assign the different components of the underlying infrastructure with their respective roles in the model. Finally, present knowledge and technology must render unexplored physiological attributes relatively tangible to enable experimental testing of the theoretical predictions. Optimally, this should result in some well developed model pertaining to a theoretical problem, rather than to a specifically tailored model of the system. The challenge naturally lies in the need to balance between two eternal defining principles, namely, the wish for reductionism, and the need to remain true to the neuronal substrate. I believe that the present state of knowledge of the basal ganglia places it quite uniquely in an optimal position to be studied precisely by this form of investigation.
In the following, I will outline my thesis from the perspective of the border between theory and practice, i.e., between models that originated in the field of artificial intelligence and the nature of the neural substrate as evidenced by electrophysiological data. I will focus on the question of acquisition and performance of adaptive behaviour in an uncertain environment comprised of numerous opportunities for satisfaction and danger. I will first summarise and combine some of the knowledge found in the vast literature on this topic from the behavioural, computational and physiological perspectives. My own studies, which resided in the junction of the three paths, will be presented in the results section, and the discussion will hopefully pave the way to future questions.

Learning and reward – behavioural implications

Until the 1970’s, the study of the behavioural aspect of learning was dominated by the stream of behaviourism in psychology. The behaviourists initially claimed that for psychology to be a true science, it must limit itself to explaining human (and animal) behaviour only by association of observable elements. This methodological assertion later turned into a theoretical claim, by which all kinds of normal and abnormal behaviour are structured through the formation of simple association between external stimuli. Despite differences in theories, behaviourists (to cite a few of the best known (Sechenov, 1996; Pavlov, 1927; Thorndike, 1911; Skinner, 1974) claim that the following basic rule give a sufficient account for learning: behaviour is followed by a consequence, and the nature of the consequence determines the tendency to repeat the same behaviour in the future. This rule is best known in its formulation by Thorndike, later coined as Thorndike’s law of effect, which reads as follows: “The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur” (Thorndike, 1911). Indeed, at least when human learning is considered, a reference is often made to the classic definition by G.A. Kimble, published during the golden age of behaviourism in psychology: ‘a relatively permanent change in response potentiality which occurs as a result of reinforced practice’ (Hilgard et al., 1961).
This definition set the basis for the field known as reinforcement learning, which has subsequently lent its name to machine learning. However, before we dive into exploration of the consequences of such learning, a semantic word of caution is warranted. As is often the case in theoretical psychology, the term ‘reinforcement’ also gives rise to a number of definitions stemming from theoretical issues in learning theory. Literally, reinforcement relates to the operation of strengthening or solidifying something. The ‘something’ strengthened is generally considered to be a learned response or the bond between that response and a stimulus (Reber, 1985). On the other hand, the same term has often been used to describe any set of circumstances that an organism finds pleasurable or satisfying. It seems that this dual use has brought forth a number of theoretical controversies that could have been avoided had this semantic overlap been acknowledged. In this text, I will refer to ‘reinforcement’ in the first sense, and the latter will be substituted by ‘reward’.

The basic principle of the reinforcement of behaviours by reward (or extinguishing them by punishment) yields a number of experimental predictions suggesting empirical measures of the strength of reinforced behaviour. Thus, it has been found that behaviours that are more likely to be rewarded, or that are followed by larger rewards are executed faster (Glimcher, 2003; Komura et al., 2005), with greater accuracy (Komura et al., 2005) and more often (Shull, 2005) than less rewarding behaviours.
A somewhat less intuitive topic is that of the impact of reward on decisions between multiple choices. One may, of course, view the experiment dealing with the effect of reward modulations on behaviour as degenerate instances of decisions, in which only one alternative is offered. However, different models of choice behaviour may only be dissociated when real decisions are encountered. Animal studies examined this question using manipulation of reward schedules. Typically, animals were confronted with the choice between two alternatives of varying reward perspectives, which were manipulated in different manners. Classical experiments on such concurrent reward schedules led to the formulation of the matching law (Herrnstein, 1961). This law states that when presented with a choice between two alternatives which produced reward with different time intervals (concurrent variable interval – variable interval schedule, VI-VI), the relative rate of response in one alternative \( R_a \) can be described as a function of the rate of reward in the same alternative \( r_a \) in the following manner:

\[
\frac{R_a}{R_a + R_b} = \frac{r_a}{r_a + r_b},
\]

where \( b \) denotes the complementary alternative (or set of alternatives).

Subsequent work extended this law to include non-linear relationships (see for example (Baum, 1979)). Still, this behaviour is not very insightful in terms of the way it is learned, since it can be shown that in a VI-VI reward schedule, behaviour that follows the matching law is optimal if full knowledge of the model is unavailable. However, in a slightly different variant of the task known as the variable ratio-variable ratio (VR-VR) schedule, in which rewards are given at each option with some probability, but are not accumulated if they not picked up, matching (now called 'probability matching') is no longer optimal. And indeed, many animals switch to the optimal behaviour of always choosing the option that yields reward with the highest probability ('maximizing') (Bailey and Mazur, 1990; Wolford et al., 2000). It was not until economists started showing interest in human gambling behaviour, that it was found that humans, (and indeed bees) exhibit the puzzling sub-optimal behaviour of probability matching (Wolford et al., 2000; Keasar et al., 2002; Vulkan, 2000). Curiously, Wolford et al report that humans with split-brain lesions perform maximizing behaviours with their right hemisphere (Wolford et al., 2000), as do
young babies (Vulkan, 2000). This seemingly counter-evolutionary phenomenon may be accounted for by the mechanism by which behaviour is acquired (see below), or by its adaptive advantage. Perhaps 'higher' species have a long enough explicit memory to encompass the complicated notion that world statistics (and among them the laws of reward) are not guaranteed to remain fixed, and it is therefore beneficial to embark on occasional exploratory expeditions to adjust their internal model of the world. One way to achieve adaptive exploration, while maintaining fairly high reward levels is probability-matching. This issue will be elaborated further in the 'machine reinforcement learning' chapter of this section and in the discussion.

Various classes of dynamic models tried to account for the behaviours that take place in the presence of reward. Despite many differences, many of them highlight the effect of received (or omitted) rewards on the changes of response probability, rather than on global calculations of optimality (Bush and Mosteller, 1955;Estes, 1961;Hinson and Staddon, 1983;Staddon et al., 1983;Staddon, 1988;Williams, 1991;Staddon and Horner, 1989). In those, response probabilities are simply governed by the law of effect. Interestingly, probability-matching behaviour is an emergent property of these models. However, one should bare in mind that such learning requires a mechanism that is not entirely trivial. A number of complications hinder straight-forward implementation of strengthening certain behaviours when reward is delivered. First, rewards are rarely immediate. Rather, reward is often delayed by long time periods with respect to the behaviours for which they are awarded. It is therefore hard to imagine that the reinforcement is achieved the reward per-se. In an action setting a long sequence of actions may transpire before reward is actually gained. This introduces the obvious problem of identifying the exact behaviour that should be reinforced. In addition, rewards are often stochastic in nature. Therefore reinforcement must be achieved through some learning of the statistics of their occurrence. Finally, different behaviours may lead to different types and amounts of reward, introducing the need for an exploratory mechanism that will reduce the danger of converging to a local minimum in the reward function. In the following chapter I will outline the ways in which the computational field of machine learning has coped with these complications.
Machine reinforcement learning

With the advance of machine learning, the concept of learning through rewarding consequences was warmly embraced by computer scientists. In a truly task-oriented frame of mind, classical machine learning theories divide the “learning” to two separate types. In the first type, the learning element, the agent, is monitored by an all-knowing element, the teacher, which informs it on the correct action to be taken in each situation. This form of learning is termed “supervised learning”, and is typically comprised of a set of examples (tasks) and their respective answers (actions). The second type is known as “unsupervised learning” because it lacks the teacher element. Such learning is usually applied to tasks aimed at uncovering regularities in input statistics. An intermediate learning approach, usually classified under the broad title of supervised learning, is called reinforcement learning. This field deals with situations in which an agent with an explicit goal acts upon the environment. The agent’s actions change the state of the environment, which in turn provides feedback (reward or punishment) on its actions. In this scheme, external reward functions as an evaluative signal, indicating the degree of appropriateness of network performance. Thus, on the one hand, reinforcement learning is a form of supervised learning, because the network receives and uses feedback information from the environment. However, this information is a scalar value, and is therefore evaluative, rather than instructive (Hertz et al., 1994). The evaluative signal is given either as a single bit of information: right or wrong, or, in the continuous case, as a value describing the degree of correctness. The correct answer itself remains unknown to the actor (unlike supervised learning). Reinforcement learning is therefore sometimes referred to as “weakly supervised learning” (Pennartz et al., 2000).

Since we aim to define the overlap between artificial intelligence and the neural system it is appropriate to examine the applicability of these types of learning in brains. Indeed, application of both the supervised and unsupervised-learning modes to modelling of brain function encounters some feasibility problems. The most obvious relates to supervised learning, since introducing an all-knowing teacher to a network is biologically unrealistic (Churchland and Sejnowski, 1992). At the other extreme, classical unsupervised learning (e.g. the original Hopfield network (Hopfield, 1982) or principal component analysis (PCA) network (Oja, 1982) is
useful for a restricted, albeit undisputedly important, set of carefully chosen problems. These shortcomings are overcome in reinforcement learning, making it an ideal substrate for tracking down its neural correlate.

To understand the general learning scheme of reinforcement learning systems, we shall introduce three additional concepts, after (Sutton and Barto, 1998): a policy describes the agent’s actions in each environmental state, i.e. a set of stimulus-response rules or associations. A value function specifies the long-term desirability of the state, based on the likely sequence of subsequent states and the rewards available in them, if a fixed policy is followed. The aim of a reinforcement learning system is to maximize the expected return, which is a function (usually discounted sum) of the subsequent reward sequence. Nearly all reinforcement learning methods are structured around estimating value functions. Given an accurate value function, extracting the optimal policy reduces to a trivial task, which we shall turn to shortly. A general solution for this would be repeated iterations of improving the estimate of the state value according to a given policy, followed by improving the policy in view of the change in the value estimate (Sutton and Barto, 1998).

Thus, the first objective of a reinforcement learning algorithm is to find an estimation of the optimal value function for each state such that the error of this estimation is zero. The typical setting for this task is that of classical conditioning, namely, stimuli (or states) that are followed by reward and the value function estimator must learn to predict upcoming reward from the state. A very influential approach to this problem was proposed by Robert Rescorla and Allan Wagner (Rescorla and Wagner, 1972), in which learning is induced by the discrepancy between what is predicted and what actually happens. However, this account for conditioning refers to a highly simplified state of affairs, where time is effectively divided into virtual slots of 'states' and 'rewards' (or in the words of the classical conditioning milieu – CSs and USs). As already noted, learning to predict natural rewards is typically hindered by the temporal credit assignment problem. To address this problem, an extension to the Rescorla-Wagner model was put forth by Richard Sutton (Sutton, 1988), which came to be known as the Temporal Difference learning model (TD learning). This learning algorithm utilizes a form of bootstrapping, in
which predictions are adjusted according to other, more accurate predictions (see description in (Sutton and Barto, 1998)).

The learning algorithm of the most basic TD learning algorithm, TD(0), is the following:

1. Estimate value of current state \( (V_t) \) as discounted sum of expected rewards:
   \[
   V_t = \sum_{k=0}^{\infty} \gamma^k r_k
   \]

2. Measure ‘truer’ value of current state: reward at present state + estimated value of next state: \( V_t = r_t + \gamma V_{t+1} \)

3. Compute \( \delta_t \), the TD error signal: \( \delta_t = V_{t+1} - V_t + r_t \)

4. Use TD error to improve the state value estimation for the next iteration:
   \[
   V_t^{n+1} = V_t^n + \eta \delta_t , \text{where } \eta \text{ is defines the learning rate.}
   \]

A classical conditioning setting is illustrated in Figure 1.1, showing the main events, the estimated value function and the TD error in two cases: received reward and omitted reward. It should be apparent from the figure and from the above equations that the TD error signal is linearly related to the state value, i.e., to the expected reward. The 0 in TD(0) refers to the fraction of the states that should be held in the memory to be modified in the next step. The parameter which is notated as \( \lambda \) (and called 'eligibility trace) ranges can range from 0, as in the basic case described above, and 1, in which the entire sequence is modified at once.
Introduction

Figure 1.1

The TD learning algorithm.

Schematic timeline of TD learning algorithm in a classical conditioning context.
a. With reward delivery
b. With omission of predicted reward
In a setting other than classical conditioning, the agent acts in order to receive rewards. Here two additional tasks should be achieved: a policy must be formed according to the estimated value, and individual choices of action should be performed at each time step. A number of extensions to the TD learning scheme of the classical conditioning setting have been proposed. In one, called the Actor/Critic method, the problem at hand is divided between two dedicated components. The critic is responsible for value estimation, achieved in the same manner as in the described above. The policy is explicitly stored in an actor element, who uses the same TD error signal to update the policy. Learning in an actor/critic paradigm is always done on-policy, i.e., only the policy that is currently employed is updated. An alternative class of algorithms do not involve explicit representation of the policy. Instead, the value function that is being estimated is that of action values rather than state values. In this way, the optimal policy emerges from comparing the values of different actions. Algorithms learning action value can employ on-policy learning (like SARSA) or off-policy learning (e.g., Q-learning (Watkins and Dayan, 1992)), which has the obvious advantage of separation between what is done and what is learnt.

Whatever the learning scheme, it is clear that in order to optimize the policy correctly, all states must be visited and all actions must be taken. This is another way to look at the third complication faced by every reinforcement system we discussed earlier. On one hand, when reward is encountered the action that preceded it will be reinforced, but on the other hand, the amount of reward should ultimately be maximized over all possible options. Thus, aside from the reinforcing values of the reward, it should also be viewed in relation to all other options, which means that all options should be tried. This notion embodies an additional key element in a good reinforcement learning strategy: exploration Vs. exploitation: the agent has to exploit what it already knows in order to obtain a reward, but it also has to explore in order to achieve better performance in the future (Sutton and Barto, 1998).

Physiological counterparts of reinforcement learning

Pioneering physiological self-stimulation studies of the neural correlates of pleasure, motivation and reward centres have identified regions midbrain regions as
mediating the sensation of pleasure and behaviour oriented towards it (OLDS and MILNER, 1954). The structures involved were believed to be the lateral septum, lateral hypothalamus and the medial forebrain bundle (MFB). It is now commonly accepted that the optimal region for self stimulation is the MFB, also known as the mesolimbic pathway, carrying dopamine from the ventral tegmental area (VTA) to the ventral striatum or nucleus accumbens (NAc). The midbrain dopamine system consists of neurons located in two nuclei, the VTA and the substantia nigra pars compacta (SNc), projecting mainly to the ventral and dorsal striati, respectively. A third pathway, from the VTA to regions in the frontal cortex is less pronounced, and probably important for other behaviours and pathologies (Lipska, 2004; Siever and Davis, 2004). The striatum serves as an input structure of the basal ganglia (see figure 1.2), a group of nuclei which forms a closed loop with the circuit, and which has been implicated with motor, cognitive and limbic roles (Haber et al., 2000). The striatum receives massive projections from diverse cortical areas, and projects in turn to subsequent basal ganglia nuclei through a number of pathways. The output nuclei of the basal ganglia project eventually to the frontal cortex via the thalamus.

![Figure 1.2 The cortico-basal ganglia network.](image)

The basal ganglia anatomy holds a number of unique properties, inspiring the study of its computational role within the brain. The first noteworthy property is numerical: The main feed-forward connectivity in the basal ganglia is funnel shaped,
with an enormous reduction of the numbers of neurons and connections in each level. In the rat, 17x10^6 cortico-striatal neurons has converge onto 17x10^6 striatal projection neurons (Zheng and Wilson, 2002). The next step entails an even larger reduction, by a factor of 100 (Oorschot, 1996) in the rat, 500 in the macaque (Percheron et al., 1987). The number of connections from the basal ganglia outputs back to the cortex via the thalamus apparently increases, although this has not been systematically quantified to the best of my knowledge. A second important feature is the high concentration of neuromodulators in the striatum. Both dorsal (caudate and putamen) and ventral (nucleus accumbens) nuclei of the striatum, which integrate the cortical information have the highest brain markers for dopamine (Bjorklund and Lindvall, 1984; Lavoie et al., 1989; Jones et al., 2001; Zhou et al., 2001; Zhou et al., 2003) and acetylcholine (Woolf, 1991; Descarrises et al., 1997; Holt et al., 1997; Zhou et al., 2001; Zhou et al., 2003), and a high amount of serotonin in the substantia nigra pars compacta (SNC), which is the origin for dopamine (Corvaja et al., 1993) and in the striatum (Van Bockstaele et al., 1993; Prado-Alcala et al., 2003). A third point to be noticed is the inhibitory nature of the connections in the main basal ganglia axis. With the exception of the connections projecting from the subthalamic nucleus (STN), all the main projections in this feed-forward network use the inhibitory neurotransmitter gamma-aminobutyric acid (GABA). Such inhibitory inter-nuclei transmission is highly uncommon in the central nervous system, but discussion of this fact is not in the scope of this dissertation.

A large majority (90-95%) of neurons in the striatum are its medium spiny projection neurons. These neurons receive excitatory glutamatergic input from the cortex (see Figure 1), and project to the globus pallidus (internal and external segments, GPi and GPe, respectively), the substantia nigra (pars reticulata, SNr, and pars compacta, SNC) and the STN with inhibitory GABAergic connections. The projections from the striatum are classically divided into two pathways (direct and indirect, see Figure 1.1), each of which exerts an opposite net effect on the target thalamic (and thus cortical) structures (Gerfen, 1992; Albin et al., 1989; Alexander and Crutcher, 1990). The remaining 5-10% of neurons are interneurons, of which 5 classes have been identified. Of major importance to my study are the giant acetylcholine (ACh) carrying interneurons, the source for the abundant ACh in the striatum. Dopamine, as mentioned, is carried by projection neurons whose bodies
reside in the midbrain structures of the ventral tegmental area (VTA), projecting primarily to the ventral striatum and the SNC. Projecting to the dorsal striatal nuclei. The synaptic anatomy of the glutamatergic, cholinergic and dopaminergic inputs in the striatum is of vast importance. It seems that in contrast to the diffuse nature of transmission usually attributed to neuromodulators in the central nervous system (CNS), dopaminergic and cholinergic synapses in the striatum are strategically positioned in very close proximity to the glutamatergic cortico-striatal synapses (Hersch et al., 1994; Yung et al., 1995; Hersch et al., 1995).

The abundance of dopamine and acetylcholine in the striatum, as well as their co-localization on glutamatergic synapses, and evidence for their reciprocal connection (Jones et al., 2001; Acquas and Di Chiara, 1999; DeBoer et al., 1996) imply a pivotal role for the interaction of the two substances in normal striatal function. This is supported by pathology: effective treatments of Parkinson's disease involve manipulation of DA and ACh in an opposing fashion, either by elevation of the extracellular level of striatal dopamine or alternatively by reduction of striatal acetylcholine (Lang and Lees, 2002; Pisani et al., 2003). These observations have led investigators of the basal ganglia to postulate the DA/ACh balance hypothesis, which states that the two transmitters act antagonistically in the striatum (Barbeau, 1962; Nisenbaum and Kitai, 1995). The fashion by which this balance is presumably achieved has been subject to much debate. Various pathways have been shown to co-exist: reciprocal inhibition of striatal acetylcholine release by the nigro-striatal DA pathway (Pisani et al., 2000; Pisani et al., 2003; Dukarch et al., 1989; DeBoer et al., 1996; Rice and Cragg, 2004) and of DA release by ACh (Kudernatsch and Sutor, 1994), in addition to their opposing effects on the excitability of striatal projection (Zhou et al., 2003) and of long term effects in the cortico-striatal synapse (Wang et al., 2006; Calabresi et al., 1998).

Dopamine and acetylcholine both play a crucial role in the control of motivation and learning. Deficit in either substance has been shown to disrupt reward-related procedural learning processes (Kitabatake et al., 2003; Matsumoto et al., 1999; Knowlton et al., 1996). Insight into the nature of the involvement of these chemicals in learning by the striatum is obtained from the analogy with TD reinforcement learning algorithm developed in the field of artificial intelligence (see
above). When presented with an unpredicted reward or with stimuli that predict reward, midbrain dopaminergic neurons (Hollerman and Schultz, 1998; Schultz et al., 1997; Waelti et al., 2001) and TANs (Apicella et al., 1998; Shimo and Hikosaka, 2001; Graybiel et al., 1994; Blazquez et al., 2002) display stereotypical responses consisting of a phasic deviation from their tonic firing rate. Congruent with the TD learning model, this response typically shifts to the earliest reward-predicting stimulus (Hollerman and Schultz, 1998; Aosaki et al., 1994; Apicella et al., 1997; Ravel et al., 2001; Shimo and Hikosaka, 2001). It would therefore seem, at first sight, that dopamine and acetylcholine released in the striatum may provide the reinforcement signal hypothesized in TD learning, enhancing each others' effects in their opposite signals.

Naturally, if dopamine and acetylcholine were to serve as teachers in the striatal system, one would expect them to exhibit teaching skills. In neurophysiology 'learning' is generally translated to synaptic plasticity, and hence 'teaching' would be attributed to inducing, or at least modulating, synaptic plasticity. Indeed, the corticostriatal synapses are known to undergo long term changes in synaptic efficacy in the form of long term potentiation (LTP) (Calabresi et al., 1998; Reynolds et al., 2001) and long term depression (LTD) (Centonze et al., 2001; Kreitzer and Malenka, 2005). It appears that both DA and ACh play a crucial role in cortico-striatal plasticity (Reynolds et al., 2001; Calabresi et al., 1998; Calabresi et al., 1999a; Calabresi et al., 2000; Centonze et al., 1999a; Centonze et al., 1999b; Centonze et al., 2001; Centonze et al., 2003). It was shown that induction of LTP in the cortico-striatal pathway is mediated by activation of dopamine D1/D5 receptors (Reynolds et al., 2001; Kerr and Wickens, 2001). Activation of M2 muscarinic ACh receptors reduces LTP and LTD at cortico-striatal synapses (Calabresi et al., 1998; Wang et al., 2006), while M1 activation enables plasticity (Calabresi et al., 1999b). It is interesting to note in this context, the connection to drug abuse. Two components need to combine for a drug to become addictive. First, it must induce a pleasurable effect for the user, and second the behaviour of drug consumption must be reinforced. These components were termed 'liking' and 'wanting' in a comprehensive review of the subject (Berridge, 1996). It stems from the arguments outlined above that the dopamine system, along with its culprit partner ACh can serve both purposes. And indeed, numerous studies have demonstrated that most psychoactive drugs of abuse activate the very same
Introduction

pathway. Thus, cocaine and amphetamines directly increase the amount of dopamine by inhibiting its reuptake into the synaptic terminals, opiate narcotics increase its release by disabling tonic inhibition on dopaminergic neurons, methamphetamine ('crystal meth') seeking behaviour is probably mediated by inactivation of ACh transmission in the ventral striatum (Hiranita et al., 2006) and the most abundant addictive drug caffeine increases cortical levels of ACh and dopamine (Acquas et al., 2002), while nicotine is an agonist of ACh receptors, probably acting through the dopamine/ACh interaction (Zhou et al., 2003; Cragg, 2006).

Throughout the years a number of models have been proposed to encompass the wide range of known properties of the basal ganglia. Since the hallmark of the best-known dysfunction of the basal ganglia, Parkinson's disease entails severe motor impairments, most models have focused on the role of the basal ganglia in the generation of action (DeLong et al., 1983; DeLong, 1990; Wichmann and DeLong, 1996). Subsequent works have incorporated other functions (or loops) subserved by the same basal-ganglia architecture (Joel and Weiner, 1994; McFarland and Haber, 2000). Ultimately, most current models highlight the main goal of the basal ganglia as that of selecting among possible competing actions (Berns and Sejnowski, 1998; Mink, 1996; Hikosaka, 1991; Hikosaka et al., 1993; Gurney et al., 2004) and habit formation and execution (Graybiel, 1995; Daw et al., 2005; Yin and Knowlton, 2006). Some of the more recent of these use the dopaminergic input as the substrate that will 'decide' which actions should be selected, and the GABAergic inter-nuclei network to execute the selection. A very appealing model uses machine reinforcement terminology to model the basal ganglia as an actor/critic network (Barto, 1995; Houk et al., 1995; Schultz et al., 1997; Bar-Gad et al., 2003; McClure et al., 2003; Gurney et al., 2004). In this family of models, the main basal ganglia axis is the actor, which is taught by the TD error provided by dopamine neurons. The dopamine neurons are part of the critic, the remainder of which varies in identity between models (Barto, 1995; O'Doherty et al., 2004).

The latter models have formed a useful framework for the study of the different basal ganglia components. In the following, we will begin by assuming the analogy of the basal ganglia to an actor/critic network, whereby the dopamine neurons and cholinergic TANs provide the critic signals, and the remainder of the network acts
according to those. We shall explore the role of DA and ACh as critics in this network. We will ask whether they can indeed serve as critics, whether they act independently or does one drive the other; whether their roles are redundant or complementary. We will find out in what way are their respective messages conveyed, and how they may be combined within a larger view of the basal ganglia.
Methods

Subjects and training

The data were collected from three macaque monkeys (*Macaca fusicularis*, two females, monkeys C and E, and one male, monkey Y), weighing 2.5–4 kg. The monkeys’ care and surgical procedures were in accordance with the NIH Guide for the Care and Use of Laboratory Animals (1996) and with the Hebrew University guidelines for the use and care of laboratory animals in research, supervised by the institutional animal care and use committee.

The monkeys were fully trained on the task and acquainted with all possible conditioned stimuli (see behavioural paradigm section below) prior to the surgery. Training consisted of familiarization with the behavioural setup (12–20 training days), followed by training on a basic version of the task in a deterministic regime in which all correctly performed trials were rewarded (60–92 training days, until a criterion of 80% correct performance was reached). For this part of the training, I used a conditioned stimulus outside the set of stimuli used in the experiment. Finally, we the set of conditioned stimuli was introduced, along with the associated reward probabilities (for 30–41 training days). The same set was then used in the recording days (45–150 days of recording from multiple structures).

Upon completion of the experiments, monkeys Y and E underwent surgery to remove the implanted recorded chamber. This procedure was performed under aseptic conditions under deep anaesthesia (isoflurane +N₂O). After removal of the chamber and surrounding dental cement (see below), the skin was sutured. Following recovery of a number of weeks the monkeys were moved to an animal shelter where they live to this day with their newly acquired friends and families.
**Behavioural setup**

In each recording and training session, the monkeys were brought from the animal facility to the recording room in a plastic primate chair, in which they were seated in the behavioural setup. During training and recording the monkeys’ non-working hand (left) was restrained in a plastic sleeve to minimize uncontrolled movements. The monkeys faced a 17" computer screen located 40 cm from the monkey, and a linear panel with 3 buttons directly beneath it. A plastic tube connected to a metal mouthpiece for liquid reward delivery was situated approximately 1 cm from the monkeys’ mouth. The screen display and reward delivery were controlled by a computer running a custom-designed behavioural program which also collected the input from the buttons that the monkeys pressed. The software I designed and programmed was tailored to control different variations of the task performed by the monkeys, as well as real-time control of execution by the experimenter. The software was written in Visual C++ (6.0) and implemented on a PC (Windows 98/2000) computer. The program was designed to communicate directly with the data acquisition system (alphaMAP, Alpha Omega Instruments, Nazareth Israel), registering all its behavioural outputs (reward, location and identity of visual stimuli) and inputs (buttons pressed and released by the monkeys). The software also contains a large number of graphic user interface menus with which the experimenter can control virtually all task parameters (timing, visual stimuli, types of trials, etc., see below). Separate training versions of the program were also supplied to assist in the different stages of animal training. All trial parameters (trial type, stimulus location and identity, and the length of the variable-duration trial segments) were randomly chosen with the same random number generator. To avoid repetition of the same random sequence, the seed of the random number generator was changed with each initiation of the behavioural program, based on absolute time.
Results I

Behavioural paradigm

The behavioural paradigm consisted of randomly ordered trials from two tasks: basic instrumental trace conditioning trials {Staddon, 2003 11230 /id} (80%–90% of trials termed reference in chapter III of the Results section) and two-armed-bandit {Vulkan, 2000 10874 /id} decision trials (10%–20% of trials, termed decision). Trials were initiated when the monkey touched the central key. After a variable delay (1.5–2.5 s in monkeys C and E, and 2–4 s in monkey Y), the visual conditioned stimulus appeared for a short period (0.3 s for monkeys C and E, and 0.45 s in monkey Y).

In the reference trials the conditioned stimulus, located either at the left side or the right side of the screen, was one of a set of four, each associated with a different probability of receiving reward upon correct trial completion (0.25, 0.50, 0.75, and 1.00). The visual stimulus occupied half of the screen (corresponding to approximately 24×36 visual degrees). The stimulus presentation was followed by a fixed hold period of 2 s (monkeys Y and C) or 1.5 s (monkey E), after which a ‘go’ signal appeared. The monkeys were required to press either the left or right key, corresponding to the location of the memorized stimulus within an allowed response time of 800 ms for monkeys C and E and 700 ms for monkey Y. Correct responses were followed (with an interval of 700 ms) by a liquid reward at the probability associated with the conditioned stimulus. No external stimulus indicated the expected time of the reward.

In decision trials, the stimulus presentation phase consisted of a simultaneous presentation of two stimuli in both possible locations, and the monkeys could choose to press either the left or right key. Equal probability stimulus pairs were not excluded. The monkeys were then rewarded according to the probability associated with the stimulus that appeared in the chosen location. All other parameters (pre-stimulus duration, stimulus duration, hold duration, maximum response time, and reward delay) were identical in decision and reference trials. All trials (incorrect, correct, rewarded, and unrewarded) were followed by a variable inter-trial interval (ITI) (3–6 s in monkeys E and C, and 5–7 s in monkey Y).
Surgery and recording

Surgical procedure

After the monkeys were trained to a behavioural criterion of approximately 80% correctly performed trials, they underwent surgery to implant the recording chamber. The surgery is designed so as to enable access to the exact stereotaxic coordinates. To this end, the optimal coronal plane is first identified using a stereotaxic atlas (Martin RF and Bowden DM, 2000; Szabo and Cowan, 1984). In our case, we tried to achieve a maximally large section of the substantia nigra, as well as large enough access to the putamen nucleus of the striatum, and the external and internal sections of the globus pallidus. We therefore aimed at the planes that ranged from A7 to A12 (anterior to the stereotaxical coronal zero, defined as the centre of the intra-aural line. After this, we establish a theoretical location for the location of the chamber (taking into account the abovementioned considerations) and calibrate the stereotaxic device to be aimed at the centre of this range.

Throughout the operation the monkeys were under isoflurane +N₂O deep anaesthesia. Vital signals (end-tidal CO₂, O₂ saturation, ECG, body temperature) were continuously monitored. All stages of the surgery are performed under aseptic conditions. They monkeys are mounted on the pre-adjusted stereotaxic device. This device is used to define the location of a 27×27 mm square piece of bone that is excised using an electrical drill. Titanium screws are implanted in adjacent skull locations and a 27mm × 27mm plastic chamber is placed over the drilled hole. The chamber was then fixed to the skull and the titanium screws using dental cement.
Results I

Localization of recording targets

We estimated the stereotaxic coordinates of the recorded structures according to MRI scans aligned with an anatomical atlas of *Macaca fascicularis* (Szabo and Cowan, 1984; Martin RF and Bowden DM, 2000). After training, a square recording chamber with a 27 mm (inner) side was attached to the skull to allow access to the basal ganglia targets. The recording chamber was tilted 40°-50° laterally in the coronal plane, with its center targeted at stereotaxic coordinates of the SNC. The chamber’s coordinates were adjusted according to MRI imaging. An MRI scan (Biospec Bruker 4.7 Tesla animal system, fast-spin echo sequence; effective TE = 80 ms and TR = 2.5 s, 13 coronal slices 1 or 2 mm wide) was performed with 150 μm diameter tungsten electrodes at accurate coordinates of the chamber. The two-dimensional MRI images were subsequently aligned with the sections of the stereotaxic atlas.

Upon reaching the target areas, as judged by the stereotaxic and MRI coordinates, the recorded cells were identified according to their physiological properties. TANs were identified by their characteristic spike shape, consisting of a long duration action potential, and large after-hyperpolarization and the typical firing pattern (Apicella et al., 1998; Aosaki et al., 1995; Shimo and Hikosaka, 2001; Raz et al., 1996). Dopamine neurons were judged by their long-duration, poly-phasic spikes, and were additionally examined for firing elevation in response to free reward (Hollerman and Schultz, 1998; Waelti et al., 2001). Neurons of the globus pallidus were identified according to their symmetric, narrow (<1 ms) and high amplitude spike shape, and the characteristic firing pattern, with most neurons exhibiting high-frequency discharge interrupted by pauses (HFD-P) and a minority with lower frequency of discharge with occasional brief high-frequency bursts (LFD-B) (DeLong, 1971). After recording, the electrode tracks were generally continued to the neighbouring structures (Nevet et al., 2004; Arkadir et al., 2004), further aiding verification of the recorded structures.
Extra-cellular recording

In each recording session, the recording chamber was filled with 4% agarose solution to constrain electrode movements due to blood pressure fluctuations. Next, a 1.65 mm (inner) diameter guiding tube is inserted through the dura matter of the monkey in the requested coordinates of the chamber using a positioning apparatus that allowed for accurate positioning in the tangential plane (MT, Alpha Omega Instruments, Nazareth Israel). The guiding tube contained eight glass-coated tungsten electrodes with an impedance of 0.2-1 MΩ at 1 kHz. The perpendicular coordinates of each electrode were determined separately by a PC controlled mechanical driving system (EPS, Alpha Omega Instruments, Nazareth, Israel) in steps of ~10-15 μm until extracellular neuronal activity of the required type was identified. Each electrode was used to record extra-cellular activity of 1-3 cells. Neuronal activity was audibly and visually monitored. After satisfactory signals were found in all (or most) electrodes, single unit action potentials were isolated using an on-line template matching algorithm (MSD, Alpha Omega Instruments, Nazareth, Israel). Then the monkeys commenced activity on the behavioural program, and recording began of behaviour, the neuronal activity and of the gaze positions. During the recording session, the experimenters continuously monitored the spike signals as detected by the MSD and graded the units according to the fractions of ‘false alarms’ (occasions when spikes were erroneously classified as belonging to this unit) and misses (when the unit emitted spikes that were not classified to it). Manual correction of the template was performed as necessary, and recorded.

Eye Movement recording

Eye movements were recorded by an infrared reflection detector (Dr. Bouis, Freiburg, Germany). The infrared signal was amplified with a gain of x500, band-pass filtered with a 1–100 Hz 4-pole Butterworth filter (MCP+, Alpha-Omega Engineering), and sampled at 750 Hz.
Data collection and analysis

Data were fed from extra-cellular electrodes to ×25 pre-amplifiers (Headstages, Alpha Omega Instruments, Nazareth Israel) positioned on the MT. Signals were then amplified to an overall gain of x10K using signal conditioning hardware (MCP+, Alpha Omega Instruments, Nazareth Israel) and then logged at 24KHz on the data acquisition system based on a micro-star board (alphaMAP, Alpha Omega Instruments, Nazareth, Israel). The electrode signals were band-pass filtered with a 300–6000 Hz 4-pole Butterworth filter and fed to online spike-sorters (MSD, Alpha Omega Instruments, Nazareth, Israel) in which spikes were detected with a template matching algorithm. These detected spikes, as well all behavioural events were logged on the alphaMAP data acquisition system PC at 12 kHz. Eye-movement data (vertical and horizontal positions of eye gaze) were sampled at 750 Hz and logged to the alphaMAP, as well as licking movements (horizontal, vertical, rotational and overall force), sampled at 375 Hz. Finally, three digital video cameras recorded the monkeys’ faces and upper and lower limbs.

Analysis of eye movements

The results described in chapter III of the Results section require special care as to confounding effects of possible differences in the gaze position of the monkeys when monkeys performed different actions in the decision task. Traces of eye movement recordings were subjected to a number of examinations to verify that future action in decision trials was not correlated with eye positions in early segments of the trial, on top of a t-test to explore systematic differences at the gaze direction the between trial types at the time of the cue presentation and at the end of the examination window.
**Principle component analysis**

To identify possible temporal patterns of eye movement during the period of interest, I performed principle component analysis (PCA) on all sequences of sampled eye positions; the analysed segments started 1 s before the conditioned stimulus display and ended 400 ms after display, at which time the examination of neuronal responses ended. In this type of analysis, the multidimensional data are searched for a smaller set of dimensions that will define a new space which will explain most of the variability in the data. These dimensions, or principle components (PCs), are ordered according to the fraction of variability for which they account. Formally, the PCs are the eigenvectors of the covariance matrix describing the data. One application of PCA is clustering of data by way of projecting the different data points on the lower-dimension space. In the Results section (Figure 3 in Results chapter III), I projected the eye positions from the reference trials in the relevant 1400 ms on the two-dimensional spaces defined by the first and second PC, and by the third and fourth PC to search for visually distinguishable clusters between the two movement directions (right and left). I then added the data points corresponding to the decision trials and visually examined their mapping in these spaces. The coefficients of the first 4 PCs in all recording sessions were also tested for differences using multivariate analysis of variance (MANOVA) (Norman and Streiner, 2000).

**Gaze index**

For a more quantitative analysis of the visual input to the monkeys prior to and during the time of the neuronal response, I devised an index indicating the amount of time the monkeys spent looking to the right. For this I established the baseline as the mean horizontal eye position of the monkey during the ITI, and the variability in this period. ‘Right’ annotates all samples in which the A/D value of the IR measurement indicating horizontal eye position exceeded the baseline by over 1 STD, and ‘left’ annotates the samples below baseline by at least 1 STD:

\[
 See_{\text{right}} = \frac{T_{\text{right}} - T_{\text{left}}}{T_{\text{trial}}}, \text{ where } T_{\text{right}} \text{ is the number of samples looking to the right in a single trial.}
\]
To transform this index to a normally distributed variable, I computed its Fisher z-transformation (Sokal and Rohlf, 1995):

$$z_{right} = \frac{1}{2} \ln \left( \frac{1 + See_{right}}{1 - See_{right}} \right)$$

and subjected this new variable to a t-test for comparison between two samples.

**Neuronal Data analysis**

Unless explicitly stated otherwise, all data analysis (of the neuronal data, as well as of eye movements and monkey performance, was conducted using custom code, written in Matlab 6.5 and Matlab 7.04 (The MathWorks, Natick, MA).

**Stability analysis**

The first step in the neuronal data analysis targeted verification of the real-time isolation quality and stability of the spiking activity. Only spike trains considered to be emitted by a single cell during real-time sorting were subjected to rate-stability analysis, in which the instantaneous firing rate of the neuron in task-neutral periods was examined for changes. For this purpose, the firing rate during the ITI period in consecutive trials in the entire recording session was graphically displayed, and the largest continuous segment of stable data was selected for further analysis (Matlab GUI developed by Dr. Yoram Ben-Shaul).

**Creation of database**

Stable cells were chosen for the task-relevant neuron database after examination for response to at least one of the behavioural events (visual stimulus, reward, and reward omission) in the reference task using a Mann-Whitney U test, $P<0.05$ after a Bonferroni adjustment to compensate for multiple comparisons. For the purpose of all results sections except section II, only cells which were stable for at least 5 trials of each condition in the reference task. Cells included in the analysis of Results chapter III fulfilled an additional requirement of at least 3 pair combinations in the decision task.
Definition of response

Cell responses to behavioural events were parameterized as the difference in average firing rate at the 400 ms following the event compared to the preceding 400 ms. The 400 ms time window was chosen as the average time in which neuronal responses in reference trials returned to baseline. Some of the results presented in chapter I of the Results section describe responses of striatal tonically active neurons (TANs). The stereotypical response of these neurons to behaviourally significant events is comprised of a series of 2-4 alterations between increases in decreases in firing. Therefore, a simple spike count in the appropriate time window cannot capture the extent of the response. I therefore pre-processed the activity of these neurons so as to obtain an absolute measure of the response. This was achieved by half-wave rectification of by means of taking the absolute value of the baseline subtracted mean firing rate at each time point. The neuronal data presented in Results section III relies on population averages of all recorded cells. Generally, for purposes of visualization the neuronal activity was represented as a peri-stimulus time histogram (PSTH). A PSTH depicts the time course of the changes in the mean firing rate as caused by an external (behavioural or other) event. The figures show a smoothed curved, filtered with a Gaussian kernel (generally with width of SD=10 ms). However, neither result in the following analysis was based on this smoothing. Rather, all analyses are based the difference in raw spike counts in the relevant time periods. These spike counts were registered separately for each trial and neuron. This way, pooling could easily be accomplished as desired: for all trials of a particular neuron with a certain behaviourally significant feature, for the entire population of neurons in the same pooling, etc. These counts were subsequently used in all the parametric and non-parametric tests performed in the results section. For the purpose of this dissertation I assume that the standard methods commonly used in analyzing neuronal data are well known to the reader. A good reference to the parametric methods and some of the non-parametric methods can be found in (Sokal and Rohlf, 1995; Norman and Streiner, 2000). The non-parametric methods used here are clearly described in (Siegel, 1956).
Results I

Neuronal correlations

Cross-correlation analysis (presented in chapter I of the Results section) characterizes the functional connectivity between pairs of neurons. This analysis shows the probability (or mean firing rate, depending on normalization) that one (‘target’) neuron will fire as a function of the time that elapsed from firing of a second (‘reference’) neuron (Perkel et al., 1967; Aertsen et al., 1989). The result is given as a continuous function of time describing the firing probability in each (positive and negative) time lag. In my analysis I used 1 ms time bins over a ±1 s time period. The result was then smoothed with a Gaussian window (σ=2 ms). For cross-correlation to carry functional significance, it should be compared with the expected cross-correlation function under the null hypothesis, i.e., that the two neurons are unconnected, and that any apparent correlation between them results from their reaction to common external events. Cross-correlation functions were therefore considered significant (with peaks, troughs or oscillations) if they differed from the shift predictor (Perkel et al., 1967) by ±0.995 confidence intervals. This analysis was performed only on spike trains emitted by cells recorded by different electrodes so as to avoid effects of spike-sorting (Bar-Gad et al., 2001).

Regression methods

In chapter II of the results section, choice behaviour is given as a function of long term dopamine response and of long term reward. To conform to the common presentation of choice behaviour and reward originally presented by Herrnstein (Herrnstein, 1961), I show the proportion of responses of one choice (arbitrarily, right) as a function of the relative dopamine response and reward delivery of the choice, compared to the alternative. To unravel the full model of interplay between the variables I employed multiple regression analysis, and partial correlation analysis (see below). It should be noted, however, that proportion data cannot serve as an independent variable in linear regression because this would violate the homoscedasticity assumption of the test. Therefore, a transformation should first be performed. For instance, the proportion variables (f) can be converted from the [0,1] range using an angular transformation (Sokal and Rohlf, 1995): \[ \theta = \arcsin(\sqrt{p}) \].
Since choices are essentially binomial variables, choice probability was also examined using multiple logistic regression analysis (Sokal and Rohlf, 1995). In this analysis the relation of continuous independent variables is related to binomial dependent variables by the following logistic model: 

\[ p = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \]

where \( p \) is the probability of the dependent variable; \( x \) is the vector of predictors and \( \beta \) is a vector of the regression coefficients. The parameters of the logistic model are estimated using maximum-likelihood estimation and thus the overall likelihood of the model can be examined and compared to other models.

**Partial correlation analysis**

A related method of describing the relevant contributions of various correlated predictors to a dependent variable is the partial correlation method. In this case one would like to differentiate between the relative contributions of two variables to a third one in a system of three highly correlated variables, but the same analysis can easily be extended to larger systems. The calculation gives the correlation between two variables when corrected for the correlation that could be expected from a common factor that both are correlated with.

Formally, the partial correlation between factors \( a \) and \( b \) given common factor \( c \) is given as:

\[
r_{ab,c} = \frac{r_{ab} - r_{ac}r_{bc}}{\sqrt{(1 - r^2_{ac})(1 - r^2_{bc})}}
\]

In this way, models of different causality schemes can be compared, similarly to path analysis (Sokal and Rohlf, 1995).
Results

Coincident but Distinct Messages of Midbrain Dopamine and Striatal Tonically Active Neurons

Summary

Midbrain dopamine and striatal tonically active neurons (TANs), presumed acetylcholine interneurons, signal behavioral significance of environmental events. Since striatal dopamine and acetylcholine affect plasticity of cortico-striatal transmission and are both crucial to learning, they may serve as teachers in the basal ganglia circuits. We recorded from both neuronal populations in monkeys performing a probabilistic instrumental conditioning task. Both neuronal types respond robustly to reward-related events. Although different events yield responses with different latencies, the responses of the two populations coincided, indicating interaction at the target level. Yet, while the dopamine neuronal response reflects a mixture between expectation and outcome in the positive domain, the TANs are invariant to reward predictability. Finally, TAN pairs are synchronized, compared to a minority of dopamine neuron pairs. We conclude that the striatal cholinergic and dopaminergic systems carry distinct messages by different means, which can be integrated differently to shape the basal ganglia responses to reward-related events.

Introduction

Striatal dopamine and acetylcholine are intertwined in anatomy, physiology, and pathology. The striatum, the primary input stage of the basal ganglia, displays the same disinhibition as the central nervous system for both dopamineergic (DA) (Loewe, 1985; Jence, 2001) and cholinergic (ACH) (Hoff, 1997) markers. Effective treatments of Parkinson’s disease involve manipulation of DA and ACH in an opposing fashion, either by several of the environmental stimuli that ameliorate symptoms. These observations have led investigators of the basal ganglia to postulate the DA/ACH balance hypothesis, which states that the two neurotransmitters act antagonistically in the striatum (Baca, 1992; Nissenbaum and Kida, 1990). This is presumably achieved by random inhibition of striatal acetylcholine release by the nigro-striatal DA pathway (Shimizu, 2003; Kishimoto, 1998; Delogu and Abbraccio, 1997) and of DA release by ACH (Kuhlenbeck and Sailer, 1994). In addition to their opposing effects on the excitability of striatal projection neurons (Otake et al., 2001), striatal acetylcholine, commonly believed to be secreted by local tonically active neurons (TANs) (Wilson et al., 1999; Bennett and Wilson, 1999; Accorsi et al., 1999), and striatal dopamine released by midbrain neurons, but also striatal dopamine in the control of motivation and learning. Deficit in either substance has been shown to disrupt reward-related procedural learning processes (Rizzolatti et al., 2003; Maffei et al., 1994; Knowlton, 1994). On the cellular level, DA and ACH play a crucial role in cortico-striatal plasticity (Tang et al., 2001; Galaburda et al., 1999; 1993; 2000; Couteel et al., 1999a, 1999b, 1999c). When presented with an unpredicted reward or with stimuli that predict reward, both dopamine and acetylcholine neurons (Bolam and Sonders, 1997; Schall et al., 1997; Wootz et al., 1995; and TANs (Baksi et al., 1998; Shimizu and Miyake, 2000; Crayton et al., 1995; Bisque et al., 2002; Dopera et al., 2002), display stereotypical responses consisting of a vibration from their firing rate. In accordance with the DA/ACH balance hypothesis, these typical responses seem to be opposite, in that the dopaminergic neurons elevate their firing, whereas the TANs firing is mainly decreased. The dopaminergic response has been recently interpreted as an early signal that inhibits the cortico-striatal system of the discrepancy between the prediction of a reward and its actual occurrence (Schultz et al., 1997; see Fodor et al., 1996; Harz et al., 2000). This hypothesis is in consonant with the computational temporal difference (CTD) model for reward and prediction of future rewards (Accorsi et al., 1994; Apicella et al., 1997; Rome et al., 2001; Shimizu and Miyake, 2000), suggesting that they may act in the same manner. A neurobiological relationship between the degree of predictability and neural responses has recently been described for midbrain DA neurons in monkeys performing a classical instrumental task (Fort et al., 2004). However, the responses of the striatal TANs were never tested in a formal probabilistic task.

Despite these similarities, it is unlikely that the information conveyed by responses of two populations affecting a single target is redundant. To examine the role of striatal TANs and DA neurons in a task involving association of environmental input to motor output, we devised a probabilistic instrumental conditioning task (Figure 1A) that enabled us to manipulate the degree of reward predictability. In such task, one of a set of 4–5 visual cues was briefly presented to monkeys in one of two possible locations on a computer screen. After a constant delay, a go signal instructed the monkey to indicate the correct location by pressing one of two keys. Correct performance was rewarded in a probabilistic manner, depending on the preceding visual cue (Figure 1B). We recorded the single unit activity of TANs in the putamen nucleus of the structure and DA neurons in the substantia nigra pars compacta (SNc) (Figure 1C) from
and technique. Thus, the experimental design allowed us to address the question of encoding of reward predictability in conditions involving the mapping of sensory information to action and at the same time to differentiate the two neuronal populations under identical conditions.

Results

Behavior

The monkeys were trained to the point that their behavioral responses were independent of trial condition, despite the fact that the different visual cues were associated with reward at different predictability levels. This enabled us to rule out differences in neuronal activity due to kinematical, behavioral, or motivational differences. This control was particularly vital in this study, since TANs have been shown to respond differentially according to the probability of behavioral response (Blazquez et al., 2002) and DA neuron responses have recently been shown to correlate with reaction time (Sahih et al., 2003) as well as due to the motor nature of the putamen (Cruncher and DeLong, 1984; Alexander and DeLong, 1985; Lee and Assad, 2003). To reduce motor variability, we restricted the monkeys’ allowed response times to 700-800 ms, which was almost at the limit of their ability. Reaction times and movement times in all recorded trials (for all monkeys) are plotted in Figures 2A and 2B, respectively. The distribution of these parameters was independent of reward probability in each of the three monkeys (p > 0.4, one-way ANOVA and Kruskal-Wallis nonparametric ANOVA).

Note, however, that the strict constraint on the response time imposed a regime of time pressure for performance of the correct response. In such cases it has been shown both theoretically and experimentally (Reddi and Carpenter, 2002; Carpenter and Williams, 1995; Rottman and Strickland, 2002; Gondek and Crooks, 2004) that performance is suboptimal, similar to the speed-accuracy tradeoff effect in motor performance. For our purposes, it was important that this reduction in performance should remain independent of trial condition. Indeed, the percentage of correct choices (Figure 2C), as well as self-aborted trials (i.e., trial break error and response omission error [pooled together in Figure 2D]), was invariant across all conditions (one-way ANOVA, p > 0.7, Kruskal-Wallis nonparametric ANOVA, p > 0.4).

Since no behavioral differences were permitted, a different measure was required to ensure that the monkeys learned the probabilities associated with the different cues and were able to utilize these to predict an upcoming reward correctly. To this end, we introduced within each recording session several probe trials (5%–10% of total trials, randomly interleaved between the single cue trials in which two visual cues were presented simultaneously at both positions. In these trials pressing either key could yield a reward whose probability depended on the key that was pressed. Figure 2E depicts the monkey’s choices. The color matrix covers all combinations presented to monkey Y. Note the clear gradient at left key preference from the upper right corner (P_{1,0} = 1.0, P_{2,0} = 9.25) to the lower left corner (P_{1,0} = 0.25, P_{2,0} = 1.0) and the lack of preference at the diagonal.

Three monkeys (Y, E, and G) performing this task. Only correct trials were used for analysis (see Experimental Procedures for a detailed description of recording sites and techniques).
Distinct Messages of Basal Ganglia Teasers

Doctoral Thesis, Geneala Morris

Results

Neuronal Responses to Behavioral Events

We recorded 192 (79, 19, and 94) DA cells and 97 (17, 68, and 15) TANs from monkeys C, D, and E, respectively, during task performance. We report only results obtained during correctly performed trials. Trials in which two visual cues were presented simultaneously were not included in the current analysis. A total of 114 DA neurons and 90 TANs showed significant responses (Mann-Whitney, p < 0.05) to at least one of the three reward-related events (visual cue, reward, and reward omission) and were selected for further analysis. Figure 3 depicts the representative responses of one TAN (Figure 3A) and one DA neuron (Figure 3B). Each row represents one type of trial, classified according to the probability that a reward would follow a correct response. All three reward-related events are distinctly represented in the activity of both neurons. However, whereas the DA neuron responds oppositely to reward versus its omission (Holtmaat and Schultz, 1998; Satoh et al., 2003), the TAN responded to these opposing events with the same polarity, although the magnitude of the response to omission was smaller. As in this example, in all cases of TAN responses to reward omission, the cross features of the reward omission response were similar to those following the visual cue and the reward.

The TAN's response to the three behavioral events differs in terms of latency, as well as magnitude of the surrounding excitation. However, within each event, the similarity between the responses for all reward probability conditions is high. In striking contrast, the DA neuron's responses (Figure 3B) clearly differentiate between the various reward probability conditions, reaching reversal of the response to reward omission (Holtmaat and Schultz, 1998). This response is in line with the previous finding obtained in classical conditioning (Fiorillo et al., 2003) and the ID signal hypothesis for DA neurons, predicting that the cue response will increase with its rewarding value (i.e., with the increase in cue-associated reward probability) and the reward response will decrease with its predictability (increase in probability).

For more elaborate quantitative analysis, cell responses to behavioral events were parameterized as the difference in average firing rate at the 400 ms following the event compared to the preceding 400 ms. For the TANs, this procedure was preceded by full-wave rectification of the response (Insels in Figure 3). A large proportion of DA neurons (102, 110, and 74) and TANs (76, 89, and 40) displayed significant changes in discharge rate and pattern following the visual cue, reward, and reward omission, respectively (Mann-Whitney, p < 0.05). Most DA neurons and approximately half of the TANs responded to all three events. The pattern shown in Figure 3, whereby DA neurons hold specific information regarding reward expectation, along with the present reality, whereby TANs provide general information regarding a potentially significant event, was consistent.
for neurons of both populations. Statistical examination of all neurons revealed that while the TAN responses were not significantly different when comparing trials of different probabilities (Kruskal-Wallis nonparametric ANOVA, p > 0.3 in all cases), the observed DA response pattern differed significantly across the different probability conditions (p < 0.001 for visual cue and reward). The differences between the responses to reward omission in the different probability conditions were found to be nonsignificant (p > 0.2). Neither type of neuron exhibited sustained change from baseline activity in response to (or preceding) any behavioral event. Neuronal activity preceding the reward did not depend on trial condition (p > 0.6), indicating that uncertainty level (Fiorillo et al., 2003) did not affect firing at this period.

Population Responses to Behavioral Events
Each of the three behavioral events (cue, reward, and reward omission) elicited responses that appeared to be characteristic of the vast majority of neuronal responses within every population, with highly similar latencies, patterns, and polarity of response, as well as a relatively constant tonic (background) firing rate. Figure 4 shows mean population responses to the visual cue (left), reward (middle), and reward omission (right) of all recorded DA neurons (Figure 4A) and TANS (Figure 4B), classified according to the reward probability. On the population level, not only did the DA neurons respond differentially to the various reward conditions, but those responses were graded in agreement with the TD hypothesis (Figure 5A) as the probability of reward increased, the cue responses became larger and the reward responses became smaller. No significant differences were observed in the reward omission responses across the different trial types. As in the case of single neurons, the population averages did not display sustained responses at any time.

In contrast to the DA case, the TANS did not respond significantly different with respect to different probabilities of reward, nor did they follow any consistent trend, either at the single neuron level or as a population (Figure 4B). To quantify this apparent difference between the DA and TANS responses, we conducted a linear regression analysis of the mean changes in firing rate during responses to the visual cue, reward, and reward omission in relation to the different reward probabilities. The results are plotted in Figure 5A, showing that the DA response to the visual cue and reward is highly correlated with the probability of reward. In sharp contrast to the TANS response, the TAN response consists of several distinct phases, which may or may not occur (initial rise, pause, and second rise), we conducted a
results. Two separate tracings of spike trains were obtained from each monkey during the 11 periods. One tracing was obtained from the contralateral hemisphere during each period. The second tracing was obtained from the ipsilateral hemisphere during each period. A computer program was used to analyze the data. The data were analyzed using a two-way analysis of variance (ANOVA) followed by a Student's t-test. The results were confirmed using a linear regression analysis. The significance level was set at p < 0.05.


during the study period. The monkeys were divided into two groups: one received
responses during each period. The first group of monkeys received no treatment during the study period. The second group of monkeys received an intracerebral infusion of a drug during the study period. The data were analyzed using a two-way analysis of variance (ANOVA) followed by a Student's t-test. The results were confirmed using a linear regression analysis. The significance level was set at p < 0.05.


during the study period. The monkeys were divided into two groups: one received
responses during each period. The first group of monkeys received no treatment during the study period. The second group of monkeys received an intracerebral infusion of a drug during the study period. The data were analyzed using a two-way analysis of variance (ANOVA) followed by a Student's t-test. The results were confirmed using a linear regression analysis. The significance level was set at p < 0.05.


during the study period. The monkeys were divided into two groups: one received
responses during each period. The first group of monkeys received no treatment during the study period. The second group of monkeys received an intracerebral infusion of a drug during the study period. The data were analyzed using a two-way analysis of variance (ANOVA) followed by a Student's t-test. The results were confirmed using a linear regression analysis. The significance level was set at p < 0.05.
and C. Foeller for sharing the method of measuring DA neurones. G. Groenink for VIP- and V. Sharp for technical assistance. The structural parts were supported by a career development grant administered by the BHF, by the United States-Israel Binational Science Foundation by the German-Israeli Binational Foundation (GT) and by the EMRF-Israel-Germany collaboration in medical research. G.M. was supported by the Hermann Stelzner.

Received: November 8, 2010
Hebrew 3rd March, 2011
Accepted: 27th July, 2011

References


Doctoral Thesis, Genea Morris


Implementation of temporal difference (TD) learning by dopamine neurons: implications for the computational algorithm

The use of temporal difference (TD) learning algorithm is appropriate only when learning is still required. In these situations, the goal of the learner is to always correctly estimate the value of each state or sensory input. This estimation must constantly be monitored, and updated when necessary in order to be able to use this information in choice of action, in allocation of resources, in planning for the future etc. In a probabilistic reward regime, as exemplified in our behavioural setup, the learning task is that of correctly learning the statistics of the reward schedule. Therefore, since each occurrence of reward is a Bernoulli experiment, the monkeys must learn to estimate the parameter of the appropriate binomial distribution associated with each of the conditioning stimuli. This estimate should vary after every trial depending on whether or not this trial was rewarded. When a reward follows a certain stimulus, the value of that stimulus (or the estimated parameter of the binomial distribution) is updated to be slightly better than has previously been estimated (except for the trivial case of p=1). Conversely, if reward fails, the estimation is updated to a lower value. The subsequent estimation of the error in estimation, or the TD error should vary accordingly: If two subsequent presentations of a certain conditioning stimulus end with reward, then the second reward must have been somewhat predicted by the first, and therefore the TD error following it should be smaller. If a CS is followed by reward, and its value is updated to a higher prediction, then an omission of reward on the subsequent trial will cause a larger discrepancy and will therefore yield negative response that is larger (in absolute value).

Accordingly, if dopamine neurons report the error in this estimation, then their activity should vary according to this instantaneous estimation. Specifically, their activity is expected to be modulated by local history. Results pointing to this conclusion have been reported in a somewhat less elaborate task (Nakahara et al., 2004). However, this expected modulation by local history depends on the specific implementation of the TD algorithm. The deviation between the TD errors of two
consecutive reward deliveries will vary depending on the susceptibility of the system to changes. Intuitively, a system that is confident in its estimates, and in which stability is a desired feature will be reluctant to change its estimates based on a single experience, and therefore this deviation would be small. If, however, accuracy and brief responses to changing environments are a valued goal, then each experience will significantly alter the estimates of values. This trade-off between stability and accuracy is encapsulated in a single parameter of the system, namely the learning rate. Formally, predictions are updated from trial to trial with a rate depending on the learning rate, $0 \leq \alpha \leq 1$:

1. $V(t, n + 1) = V(t, n) + \alpha \cdot \delta(t, n)$
2. $\delta(t, n) = r(t, n) + \gamma \cdot V(t + 1, n) - V(t, n)$

where: $V(t, n)$ denotes the value function of time t in trial n, $\delta(t, n)$ denotes the temporal difference error in time t of the same trial, $\gamma$ is a discount factor for future rewards and $r$ denotes reward.

In the following analysis I investigate the effect of the local history (previous reward) on the size of the dopamine responses to the conditioning stimulus (CS), reward and reward omission.

**Results**

*Population response*

For each neuron, the trials were separated according to the outcome of the previous trial in which the same CS was presented. Peri-stimulus time histograms (PSTHs) were calculated separately for both groups of trials. Cells were subjected to stability analysis, but were not rejected due to insufficient numbers of trials (in order to maximize the size of the database). The average population responses ($n=113$ neurons) for the various events in the $p=0.75$ trials are plotted in Figure 3.1(a). Note that due to the probabilistic nature of our task the number of trials included in each average varies systematically with trial condition: only 25% of the trials belong to the 'previous disappointment' condition, and 6.25% correspond to successive disappointments. The level of noise is, therefore, expectedly different between the groups.
As is evident from the plots, the responses of the population did not show a local history effect. The same impression emerges from considering all other conditioning stimuli (Figure 3.1b,c).

Figure 3.1 **History dependence of responses.**
Responses to conditioning stimulus (left), reward (centre) and reward omission (right) separated according to receipt of reward in the previous trial. (a). trials with P=0.75 (b) trials with P=0.5 (c) trials with P=0.25. red traces represent previously rewarded trials and black traces – previously unrewarded trials.
To quantify this apparent lack of effect of previous history, we subjected the responses to unbalanced 2-way ANOVA. The analysis revealed a single, main effect of probability in the 'CS' and 'reward' events (p<0.0001 in both cases, df=2,1). This effect was not found in the disappointment and reward-time conditions (p>0.3). This is consistent with (Morris et al., 2004). By contrast, neither main effect of previous trial was observed (p>0.3) nor interaction effect were found.

**Single cell responses**

It may be argued that subtle differences on the single cell level are averaged out in the population mean. To examine the dependence of the cells' responses on reward history, the response of each cell was quantified as the mean change in firing rate during the 400 ms following each event, as compared to the preceding period.

Figure 3.2 shows the responses of each cell to each of the events following reward, plotted against the responses following reward omission. The identity line is shown for comparison.

**Figure 3.2 History dependence of responses – single cell responses.**
Scatter plots of responses to conditioning stimulus (red), reward (green) and reward omission (blue). Responses in trials when previous trials were rewarded are plotted against those in which reward was not received. Black line indicates identity of responses in both conditions.
The responses of the neurons to the different events in both contexts were subjected to a paired $t$-test. The only significant context effect at the $\alpha=0.05$ significance level was that of the cue in the $p=0.25$ condition. This result was based on 89 neurons, but on a very low number of trials for each neuron. When a criterion of minimum 5 trials per neuron was employed this effect was found to be insignificant, based on $n=14$ observations. It should be noted that we did not enough neurons that met this criterion to check the effect on the reward response.

These results, in which the actual occurrence of reward in the preceding same-CS trial bares no effect on the dopamine response, are different from those reported in (Nakahara et al., 2004). One methodological difference is their use of different measures of the response. They compared the 100-500 ms after reward. To verify that our quantification methods did not cause this difference we used the same measure, but this did not alter our results. Alternatively, this discrepancy could point to true differences in learning rate imposed by the behavioural differences that our tasks imposed. Two related differences in our behavioural paradigms come to mind: first, the difference in task complexity: instrumental conditioning, as done in our experiment, is more difficult to learn than classical conditioning. This leads to a second, potentially important difference: It is very probable that our monkeys trained on this task for a much longer period before recording. This could mean that the learning rate in our case was initially much lower (due to difficulty of the task), or that once a stimulus is learned (in the 'over-trained case') the learning rate goes down to 0. Whatever the reason, since in our case the learning rate seems to be effectively 0, then if the account of Niv et al (Niv et al., 2005) to the ramp found in (Fiorillo et al., 2003) is correct, our data should not produce a comparable ramp. Indeed, in our hands dopamine neurons do not show ramping activity from the time of CS presentation to the time of expected reward. We would then predict that if a similar analysis to what is reported here should be conducted by Fiorillo and coworkers, it should yield a larger learning rate.

When would we expect to find the effect of the previous trial's outcome? This depends on the manner by which the updating of the value function propagates back in time. This relates to a third constant, $\lambda$. With $\lambda=0$, TD(0), a prediction error at the
time of reward in each trial will be apparent in the following trial only at the immediately preceding time step. This error will back-propagate to the time of CS within T trials, T being the number of time steps between the CS and the reward. When λ=1, the values of all the time steps in the trial will be effectively changed in the next trial. For any other λ, assuming that each trial contains more than just 2 time steps, that of (one representing the CS and the other - the reward), we should not see a difference at the time of the CS depending on the previous trial. This is because in our case the expectance is not context dependent (Nakahara et al., 2004), rather, in each trial the rewards were drawn independently from a binomial distribution with the parameter p. Note that this also differs from (Fiorillo et al., 2003), since the 'probabilities' were accurate for blocks of 8 trials with the same CS.
Midbrain dopamine neurons encode decisions for future action

Genaca Morris1,2, Alon Nevet3, David Arkadir2, Eilon Vaadia1,2 & Hagai Bergman1,2

Current models of the basal ganglia and dopamine neurons emphasize their role in reinforcement learning. However, the role of dopamine neurons in decision making is still unclear. We recorded from dopamine neurons in monkeys engaged in two types of task: reference trials in an instructed-choice task and decision trials in a two-armed bandit decision task. We show that the activity of dopamine neurons in the decision setting is modulated according to the value of the upcoming action. Moreover, analysis of the probability matching strategy in the decision trials revealed that the dopamine population activity and not the reward during reference trials determines choice behavior. Because dopamine neurons do not have spatial or motor properties, we conclude that immediate decisions are likely to be generated elsewhere and conveyed to the dopamine neurons, which play a role in shaping long-term decision policy through dynamic modulation of the efficacy of basal ganglia synapses.

The art of associating sensory information with appropriate behavior or decision making has been investigated through the prism of a multitude of fields. The search for psychological and neural correlates of decision making was paralleled by machine learning research. One form of machine learning, reinforcement learning, has achieved popularity because of its efficiency and its resemblance to real-life situations. Developments in reinforcement learning have led to a powerful learning algorithm known as temporal difference (TD) learning. TD learning, originally used for modeling classical conditioning, is based on evaluating sensory inputs, or states, by assigning them a value according to the anticipated reward. Learning to optimize this evaluation is achieved by constant comparison of the value of the current state with its previous estimation. When a discrepancy arises, this difference, termed the TD error, is used to improve estimation of the state value.

The classical conditioning context provides an inadequate description of the typical reinforcement learning setting in which agents act upon sensory information to execute behavioral decisions. Moving from passive TD learning to active control requires modification of the computational algorithm, as the aim of learning has now shifted to optimization of actions in different states to maximize the long-term accumulated reward. The actions affect not only reward, but also the transition from one state to another, an outcome that must also be learned. This challenge is resolved by reinforcement learning models that incorporate actions into different variations of TD algorithms using the TD error to update the state evaluation and to adjust the set of rules that govern the decisions in each state, or the policy. Policy optimization can be achieved in a number of fashions. One way is through the design of specialized actor/critic network architecture. In these networks, the TD error is used to teach two separate elements, which, when combined, result in efficient action selection. The critic estimates the values of all encountered states (as in the classical conditioning context), whereas the actor stores the policy and performs actions. Each action can lead to a different state. This may cause a deviation from the estimated value of the previous state. The resulting change, the TD error, is fed back to the actor by the critic and is used to shape the desired policy. Another alternative class of algorithms does not involve explicit representation of the policy but relies on direct assessment of the value of state-action pairs (also termed action values or Q values) rather than the value of the state alone. Thus, both the actor-critic and Q-value estimation models are taught by a TD error. This error signal is independent of the action in the actor-critic architecture, whereas a Q-value error signal is affected by the chosen action.

The phasic response of midbrain dopamine neurons located in the substantia nigra pars compacta (SNC) and the ventral tegmental area (VTA) is a likely neural correlate of the TD error, thus underscoring the applicability of the TD learning algorithm to neural learning. Because the basal ganglia network, the main target of dopamine innervation, is commonly regarded as an action selection and generation system, the dopamine signal is incorporated as the critique in actor/critic TD models of the basal ganglia. In these models, the dopamine signal is used to reinforce behavior by adjusting synaptic efficacy in the appropriate neural circuits of the input layer of the basal ganglia networks—that is, the striatum. Some models also use the dopamine signal to directly select possible actions. Electrophysiological recordings of dopamine neurons typically involved classical conditioning or instructed-choice instrumental conditioning tasks, but the role of dopamine in behavioral decisions involving competing actions and in policy formation has not been explored experimentally.
ARTICLES

RESULTS

Reference dopamine responses determine decision policy
We observed no differences in motor parameters (reaction time, movement time) between the two trial types (t-test, P > 0.2). Thus it can be assumed that the monkeys used similar motor strategies and could use the information they gathered in the reference trials to determine their behavior in the decision context. Two related parameters in the reference trials could impact the decision policy: the delivered reward and the TD error–like dopamine activity in these trials. We examined the monkeys’ choices in the decision trials (C), and their relation to the reward rate (R) and dopamine responses to the conditioning stimuli (Q) in the reference trials (Fig. 1). Because dopamine activity was highly correlated with reward rate ($R^2 = 0.91$), the policy-determining factor could be the reward rate itself or the dopamine activity. In one model (Fig. 2a, inset), decision choices are governed by the reference reward rates, which independently also modulate the reference dopamine activity. In the alternative model (Fig. 2d, inset), the impact of reference reward on decision choices is mediated by the reference dopamine activity.

Reference reward rates (Fig. 2a) were computed as the a priori reward probability corrected by the monkeys’ error rates on those trials. In the reference trials, the reward-choice relationship was monotonic (Fig. 2b). The monkeys’ policy was thus a suboptimal probability matching strategy (Fig. 2e) ($R^2 = 0.884, P < 0.001$).

$$C_{\text{right}} = C_{\text{left}} - R_{\text{right}}$$

where $C$ is the probability of a particular choice and $R$ is the probability of being rewarded on that choice. Logistic regression analysis, which should be applied to relationships between a proportion and a continuous variable, yielded a highly significant relation (likelihood ratio test, $P < 0.001$).

Comparable studies in human subjects report a similar monotonic relationship in inexperienced gamblers, in contrast to trained gamblers who tend to maximize their return. Our design, in which decision trials were only sparsely embedded in the reference trials, is a prime model for this situation. Recent studies in repeated decision tasks show local dependence of choice behavior on reward history. Our task...
Decision choices are not predicted by early gaze shifts. Before examining dopamine activity during decision trials, we must rule out possible confounding effects of different gaze positions before and during the neuronal response. We compared the horizontal eye positions recorded during the later part of the 'start' period with those during the conditional stimulus presentation (Fig. 1). We separated, according to future action, the sets of the horizontal axis of eye positions recorded during reference and decision trials (Fig. 3a). In this example session, the eye positions in the reference trials at the time of stimulus presentation (indicated by arrowhead) differed slightly according to stimulus position (and, consequently, according to the direction of the future arm movement), but in decision trials, the eye positions were similar regardless of future movement.

To quantify possible differences in the visual inputs to the monkeys (and to SNc dopamine neurons) , which may have affected the neuronal results, we first examined differences in gaze direction in the decision trials for trials in which opposing actions were taken. A two-tailed t-test between the groups of eye positions at two time points—the time of stimulus presentation and 400 ms after presentation (neuronal responses were examined in this 400-ms window)—indicated no differences (P > 0.3 in all recorded sessions). We further examined the gaze positions by principal component analysis (PCA; Methods) by taking the 1 s preceding and the 400 ms following the visual stimulus presentation (green line in Fig. 3a) and projecting all traces on the space defined by the first and second principal components (Fig. 3b, left) and by the third and fourth components (right). In this example, the first four components explain 77.4% of the variability. In the reference task, the two movements were reflected in the gaze positions, but the projections of the eye movements in the decision tasks overlapped, indicating that they did not depend on future movement. We repeated this analysis on all recording sessions. In all cases, the first four components accounted for >70% of the variability. In no session were the decision trials separable based on future movement. As expected, the separation between clusters in the period surrounding the 'go' signal was far better both in reference and decision trials. Finally, for each decision
ARTICLES

Figure 4 Dopamine neurons code the TD error of action values. (a) Dopamine responses to pairs of conditioned stimuli in decision trials, presented in ascending order according to state value (defined as the average probability of reward following each pair). Bars reflect mean; error bars are S.E.M. (n = 37 neurons). (b) Dopamine responses to pairs of conditioned stimuli in decision trials, separated according to the chosen action. Bars reflect mean; error bars are S.E.M. (n = 37 neurons). (c) Dopamine responses to conditioned stimuli in reference (empty circles) and decision (filled circles) trials, as a function of the action value (defined as the average probability of reward following each action). In the reference trials, the expected reward probability is corrected for response errors. By definition, there are no response errors in the decision trials. Points reflect mean ± S.E.M. (n = 37 neurons). (d-f) Dopamine responses to reward delivery in the rewarded decision trials. All conventions are as in a-c.

Decision dopamine responses reflect future action

The monkeys' probability matching policy, which exhibits a variability of responses in identical situations, allowed us to explore dopamine neuron activity during decision making. The activity of different neurons was pooled to analyze the dopamine population response. The dopamine response in the reference trials accurately represents the TD error in the estimation of the state value—that is, the average expected reward. On the other hand, in decision trials responses of the population of dopamine neurons to each of the decision stimulus pairs were pooled according to their state value, these responses were statistically indistinguishable (P > 0.1, one-way analysis of variance (ANOVA); Fig. 4a). However, in contrast to the instructed-choice reference trials, the prospects of reward in the decision trial depended on the monkeys' choice of future action. We therefore separated the decision trials based on the action taken at the completion of the trials. When the mean population responses to the presentation of six pairs of nonidentical stimuli were separated according to choice of future action, in all pairs the dopamine signal was significantly higher when the monkeys chose the key associated with the higher probability of reward (post-hoc comparison, P < 0.01; Fig. 4b).

To uncover the effects underlying this variability, a two-way ANOVA was applied to the data. This analysis showed a clear main effect of chosen cue (F < 0.001), with no effect of the discarding cue (P > 0.4), and a marginal interaction effect (P = 0.05), resulting from the difference in dopamine responses when the 75% cue was chosen. We therefore pooled the data of the dopamine responses to the conditioned stimulus in the decision trials according to the value of the chosen actions (Fig. 4c, solid circles). Response of the dopamine neurons to the conditioned stimuli differed significantly according to future actions (one-way ANOVA, P < 0.001). Furthermore, the responses were proportional to the theoretical value of the different actions, similarly to the differential responses in the reference trials.

This finding points to a new interpretation of the dopamine responses observed in previous classical conditioning and instructed-choice experiments. Under these conditions, action is trivial or non-existent and therefore does not alter the neuronal responses. To
illustrate the full dependence of the dopamine response on the expected reward probability given the chosen action (that is, the state action value). We combined the results from the reference (empty circles) and decision trials (linear regression, $R^2 = 0.972$; Fig. 4c). To examine the consistency of the dopamine responses with the TD error signal associated with the action value (Q value), we analyzed the complementary dopamine responses to received reward in the decision trials (Fig. 4d–f). As predicted, the dependence of dopamine responses on the action value was reversed but see Methods for statistical limits of this analysis). The responses of all neurons were consistent with this picture (Supplementary Fig. 2 online).

Finally, we examined the time course of development of choice-related activity in the dopamine neurons (Fig. 5). The similarity of the time course of neuronal activation in the decision and reference trials (Fig. 5a). Further, for the alternative directions, dopamine neurons code the action value in two stages, the first relating to the state value and the second adjusting for the action value. Choice-related activity was defined as the time course of dopamine neurons in response to the high- versus low-probability cue choices in the decision task (Fig. 5b). The dopamine response crossed the upper limit of the 95% confidence interval 112 ms after stimulus presentation. This time course is very similar to the development of information regarding the state value in reference conditions in6. Therefore, although this differentiation may serve as an indication of the upper bound on the time of decision formation, it cannot provide us with further insight into the dynamics of the decision process.

DISCUSSION

The results presented in this report provide key insight into the role of dopamine in decisions, both in the long-term modification of behavior and in immediate decision making. First, probability matching decision behavior is likely to be mediated by the activity of dopamine neurons rather than by reward in the reference trials, suggesting that the history of dopamine responses shapes long-term behavioral policy. Second, the activity of dopamine neurons reflects future choice of action as early as 112 ms after the presentation of the conditioning stimulus. This has implications for the position of dopamine in the hierarchy of decision making that is, it is likely that dopamine neurons receive information about the decision from another structure. These results also hold with stimuli from previous lesion and behavioral studies regarding the long-term rather than immediate effect of dopamine on reward-oriented behavior.

An alternative conclusion would be that the signal of the dopamine neurons reports the error for the state value and that it directly determines decision. Some models have adopted this view, in which the TD state value signal has an additional effect on the probability of an upcoming action. However, because the dopamine signal is extremely homogeneous in the origin structures and highly widespread at the targets, the possibility of separate TD signals for each alternative can be excluded. Therefore, such models of the direct effect of dopamine on immediate decision can only apply to simple situations involving evaluation of a single alternative. In our decision setup, as in many real-life situations, this is not the case. Therefore, if the dopamine signal is used for immediate decision, the multiple choices must be translated into a single choice. For example, the deciding circuits may use the dopamine TD signal to determine the probability of the action that maximizes reward expectation. However, this strategy will collapse in scenarios where there are multiple (>2) choices.

We therefore favor the complementary hypothesis, that dopamine neurons are already informed of an upcoming action. The striatal projection neurons could follow a probabilistic policy, which is shaped by the history of dopamine reinforcement. Two broad classes of reinforcement learning models, incorporating an action generation mechanism into different variations of TD algorithms, were proposed to account for decision making. In the actor/critic model, the network includes a separate policy-performing element, the action, as well as a reward predicting element, the critic. In the other model, the behavior is integrated into the evaluation process, assigning a separate value (Q value) to each possible behavioral choice in every state. Most reinforcement learning models of the basal ganglia adopt the actor/critic view and thus stress the importance of learning state values. This model was supported by single-unit recordings of dopamine neurons and by other studies using classical conditioning and instructed-choice designs, but has never been tested in decision contexts. Our results call for a reappraisal of the current computational models of dopamine and the basal ganglia so that they incorporate the learning and estimation of Q values (as achieved by the SARSA learning algorithms or advantage learning) into the learning and decision algorithms of these neuronal structures.

METHODS

Animal training and behavioral tasks. Data were obtained from three macaque monkeys (Macaque fascicularis, two females, monkeys C and E; one male, monkey Y) weighing 2.5–4 kg. Care and surgical procedures were performed in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals (1996) and with Hebrew University guidelines for the care and use of laboratory animals in research, supervised by the institutional animal care and use committee.

The monkeys were trained on a task with randomly ordered, appetitive, instructed-choice, instrumental conditioning trials (reference, 800–900% of trials) and two-armed bandit decision trials (50–10% of trials) (Fig. 1). In all trials, the monkeys faced a 17” computer screen placed at a distance of approximately 40 cm. A panel with three keys was located at arm’s length. Trials were initiated when the monkey touched the central key. After a variable delay (1.5–5 s in monkeys C and E, 2–4 s in monkey Y), the visual stimulus condition appeared for a short period (0.3 s for monkeys C and E, 0.5 s in monkey Y).

In reference trials, the conditioned stimulus, located either on the left or the right side of the screen, was one of a set of four, each associated with a different probability (0.05, 0.5, 0.75, and 0.95) of receiving an equal amount of reward upon correct trial completion. The visual stimulus occupied half of the screen (corresponding to approximately 24 × 36 visual degrees). The stimulus presentation was followed by a fixed hold period (2 s for monkeys C and E; 1.5 s for monkey Y), after which a go signal appeared. The monkeys were required to press either the left or the right key, corresponding to the location of the remembered stimulus, within an allowed response time range of 800 ms for monkeys C and E and 700 ms for monkey Y. Correct responses were followed (with an interval of 100 ms) by a liquid reward, according to the probability associated with the conditioned stimulus. No external stimulus indicated the expected time of the reward.

In decision trials, the stimulus presentation phase consisted of a simultaneous display of two stimuli in both possible locations, and the monkeys could choose to press either the left or the right key. Equal-probability stimulus pairs were not excluded. The monkeys were then rewarded according to the probability associated with the stimulus displayed in the second location. All other parameters (prestimulus duration, stimulus duration, hold duration, maximum response time and reward delay) were identical in the decision and reference trials.

All trials (incorrect, correct, rewarded and unrewarded) were followed by a variable intertrial interval (ITI): 3–6 s in monkeys C and E, 5–7 s in monkey Y. The monkeys performed 500–500 trials on most training and recording days. Monkeys were trained for 5–6 weeks and were allowed free access to food and water on the weekends. The monkeys were not trained for gaze fixation.

The trial sequence was completely randomized. Every trial was chosen with a preset probability, to be either a reference trial ($p = 0.8$ or $0.9$) or a decision trial ($p = 0.1$ or 0.3). In reference trials, the right or left locations of the
ARTICLES

Results III

Mann-Whitney U-test (P < 0.05 after a Bonferroni correction for multiple comparisons). Only cells that were stable for at least five trials of each condition in the reference task and at least three pair combinations in the decision task were included.

Cell responses to behavioral events for ANOVA, linear regression, logistic regression, and partial correlation analysis were parameterized as the difference between average firing rate in the 400 ms following the event and that in the preceding 400 ms. The 400-ms time window was chosen as the average time in which neuronal responses in reference trials returned to baseline (Fig. 5). Some individual neurons had longer responses. We therefore repeated the analyses with 500-ms windows. This analysis yielded comparable results. The neuronal data presented here rely on population averages of all recorded cells, normalized by the number of trials recorded from each cell. Averaging responses normalized for baseline firing rate, as well as for maximum response rate, did not affect the results.

For linear regression of the relative fraction of choice as a function of the relative neuronal and reward reinforcer, the proportion variables (p) were converted from the [0, 1] range using an angular transformation:

\[ \theta = \arccos(p) \]

Choice probability was also examined using logistic regression analysis, which investigates the relationship of continuous independent variables to binomial dependent variables by the following logistic model:

\[ P = \frac{e^{\theta \cdot x}}{1 + e^{\theta \cdot x}} \]

where \( P \) is the probability of the dependent variable, \( x \) is the vector of predictors and \( \theta \) is a vector of the regression coefficients. The parameters of the logistic model were estimated using maximum likelihood estimation, and the overall likelihood of the model can be examined and compared to other models.

Statistical analysis of the neuronal data during reward delivery in decision trials is problematic, owing to the experimental design and the behavioral strategy of the monkeys. Because of the probabilistic nature of our reward schedule, the number of rewards increased consistently with the dependent variable, the action variable, creating a bias in the respective sample sizes and their variability. This effect was also inflated by the monkeys' probability matching behavior (Fig. 2B).

Traces of eye position recordings were subjected to three statistical tests to verify that future action in decision trials was not correlated with eye position in early segments of the trial. A t-test checked for systematic differences in gaze direction between trials (at the time of the stimulus presentation and at the end of the examination window of neuronal responses (400 ms after onset). To identify possible temporal patterns of eye movement, PCA was conducted on all sequences of sampled eye positions the analyzed segments started 1 s before the stimulus display and ended 400 ms after display, at which time our examination of neuronal responses ended. In this type of analysis, the multidimensional data was searched for a smaller set of dimensions that define a new space that explains most of the variability in the data. These dimensions or principal components are ordered according to the fraction of variability for which they account. Formally, the principal components are the eigenvectors of the covariance matrix describing the data. One application of PCA is clustering of data by projecting the different data points onto the lower dimension space. We projected the eye positions from the reference trials in the relevant 1,400-ms trial segment onto the two-dimensional spaces defined by the first and second components and by the third and fourth components (Fig. 1B), to search for visually distinguishable clusters between the two movement directions (right and left). We then added the data points corresponding to the decision trials and visually examined their mapping in these spaces. Finally, we devised an index indicating the amount of time the gaze was directed toward the right during segments starting 1 s before stimulus display and ending 400 ms after display. We established the baseline as the mean horizontal eye position during the ITI. Variability was assessed in the same period. Right' denotes all samples in which the AOD value exceeded baseline by over 1 s.d., and 'left' denotes the samples below baseline by at least 1 s.d. (more conservative thresholds were also checked, yielding qualitatively similar results):

\[ \text{Score} = \frac{T_{\text{right}} - T_{\text{left}}}{N_{\text{total}}} \]

where \( T_{\text{right}} \) is the number of samples going right in a trial.
To transform this index into a normally distributed variable we computed its Fisher z-transformation:[4]

\[ Z_{\text{age}} = \frac{1}{2} \ln \left( 1 + \frac{S}{S_{\text{age}}} \right) \]

and subjected this new variable to a t-test.

All data analysis was performed using Matlab 7.0 (MathWorks) code.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank Y. Pedra, Y. Sako, Y. Rikitake, M. Nishida, Y. Nishida and R. Probst for fruitful discussions and comments on earlier versions of this manuscript and G. Gradman and V. Novak for technical assistance. This study was partly supported by a Center of Excellence grant administered by the Israel Science Foundation (ISF) and by a Fighting Parkinson grant administered by the Netherlands Foundation of the Hebrew University (HUNA). GM was supported by a Horizons fellowship.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience

34. Matsui, R. & Ueda, M. Prefrontal-prefrontal and prefrontal-prefrontal connectivity and structure of the middle temporal area (Euler Science, Amsterdam, 2004).
Supplementary Information

Dependence of choice upon local reward history

Recent primate studies of decision making have examined decisions made by monkeys in relation to temporally local reward history (Lau and Glimcher, 2005; Sugrue et al., 2004). We examined our behavioral data for similar effects by means of choice triggered averaging (CTA) of the local reward history. In the CTA analysis, for each choice in the decision trials we computed the sequence of probabilities of being rewarded in each of the preceding 40 reference and decision trials consisting of the same (chosen or un-chosen) stimulus. This analysis was conducted separately for the chosen and un-chosen stimuli. We then examined the choices made by the animals for differences in the resulting CTA functions. The results are illustrated in Figure S1. Clearly, no effect of local history can be observed in either CTA, as indicated by the lack of effect along each CTA curve. Moreover, in each of the decision trial types, the CTA functions were independent of the choices made by the monkey, and accurately reflected the pre-assigned reward probability, as indicated by the similarity between the pairs of iso-choice CTA curves in Figure S1.

This result, in which the monkeys’ choices reflect global probability-matching, rather than reliance on a local computation of reward probability, is intuitively expected from our task design. By contrast to the experiments reported by (Lau and Glimcher, 2005; Sugrue et al., 2004), where all trials involve choices between two alternatives, decision trials occurred very seldom in our task. Moreover, the monkeys were faced with 10 possible decision pairs; therefore, two decision trials involving a particular pair of stimuli were typically separated by ~50 other reference and decision trials. Furthermore, for any given decision trial, the most recent reference trial with one of the relevant stimuli is removed by four trials on average, and therefore the length of the sequence of six recent presentations of each of the two stimuli is effectively ~25 trials. In addition, because the monkeys could not anticipate which pair of non-identical stimuli would be presented in the next decision trial, they would have to keep traces of all four such sequences at all times to use reward-history information. Finally, our monkeys did not face repeated changes in the reward
probabilities assigned to each stimulus; therefore, there would be no apparent need for
the monkeys to exercise such local averaging of reward probability, which, as
demonstrated, would take an unreasonable toll on their memory.

Contribution of dopamine and reward rate to decision policy: partial correlation

Dopamine responses to conditioned stimuli in the reference task are highly
correlated with the obtained reward in the same trials. Therefore, it is extremely
difficult to differentiate between the relative contributions of the dopamine responses
vs. the obtained reward in the reference task to the behavioural policy in the decision
task. In addition to fitting a logistic model to the data (see main text), the two possible
modes of interplay between these variables (insets in Figure 2b,d) can be
distinguished by examination of the corresponding partial-correlation values. In the
model depicted on the left-hand inset to Figure 2b, the correlation between decision
choice behaviour and dopamine reinforcement in the reference task is a by-product of
reward rate in the reference task, which is the common factor. Therefore, one would
expect the partial correlation coefficient rCD.R, which eliminates the reward effect
(R) while describing the correlation between choice behaviour (C) and dopamine
reinforcement (D), to be null. Conversely, in the alternative model (Figure 2d, inset),
which places the dopamine response as the mediator between reward and choice
behaviour, rCD.R should be significant. If the entire effect is mediated by dopamine,
then rCR.D should also be null.

Calculation of the partial correlations revealed that rCR.D=0.21, which is not
significantly different from zero (P>0.4). The partial correlation of choice behaviour
with dopamine reinforcement, eliminating the reward effect, was rCD.R=0.65,
(P<0.01), indicating that dopamine indeed mediates the choice behaviour regardless
of reward, in support of the model in Figure 2d. Note that in order to overcome the
problems imposed by the 0-1 boundaries on probability values, we used angular
transformation of the choice probability (Sokal and Rohlf, 1981), and estimated the fit
of the linear regression and the partial correlations (R^2, P values) on the transformed
data. The data in the manuscript Fig. 2 are re-transformed to probability value for
illustration purposes only.
Figure S1 Choice triggered averaging (CTA) functions in decision trials. Each panel describes one decision pair, as indicated on top. Each trace relates to a particular choice and one of the two conditioned stimuli (as indicated by the color code). The CTA function represents the sequence of probabilities of being rewarded in the 40 preceding reference and decision trials with the respective conditioned stimulus.

$$CTA_{ij}(t) = \frac{1}{N_j + 1} \sum_{j'=0}^{N_j} R_{i-j'}$$  

where: $i$ denotes a choice, $j$ denotes a conditioned stimulus, and $j' \in \{0..n\}$ is the serial number of a preceding trial with conditioned stimulus $j$. $t = 0..40$ denotes time (in trials).
Figure S2 Conditioning stimulus responses according to choice value. Distributions of the responses contributing to figure 4c of the main text. Binsize = 1.22 spikes/s
Physiological studies of information processing in the normal and Parkinsonian basal ganglia: pallidal activity in Go/No-Go task and following MPTP treatment

Genela Morris1,2, Yaron Hershkovitz1, Ayal Raz, Alon Nevet and Hagai Bergman1,2,3,*

1 Department of Physiology, Hadassah Medical School, The Hebrew University, Jerusalem, Israel
2 Center for Neural Computation, The Hebrew University, Jerusalem, Israel
3 Eric Rudoal Center for Neurodegenerative Diseases, The Hebrew University, Jerusalem, Israel

Abstract: Understanding the role of the basal ganglia in day to day behavior is critical for a better understanding of the role of these structures in pathological states — such as Parkinson's disease. To elucidate this connection, we studied pallidal activity in a monkey performing a delayed release Go/No-Go task and in monkeys treated with the dopaminergic neurotoxin — MPTP. We compared the results with the predictions of the action selection and reinforcement driven dimensionality reduction models of the basal ganglia. The fraction of responding pallidal neurons, as well as the ratio of positive to negative responses, were equal in the Go and the No-Go modes. The fraction of pallidal neurons with significant responses following the trigger signal (19/26) was higher than that following the visual cue (11/26). However, the fraction of negative responses was significantly higher following the cue signal (47%) than that following the trigger signal (22%). Most (60%) of the cue responses were sensitive to the laterality of the cue, whereas only 25% of the responses following the trigger signal were sensitive to the cue or movement direction. Finally, pallidal spiking activity was not correlated in the normal behaving monkey, and became highly synchronized following MPTP treatment. We conclude that pallidal activity in the normal monkey is consistent with the model of action selection, assuming that action is selected following the visual cue. However, the reinforcement driven dimensionality reduction model is consistent with both the Go/No Go responses and the normal/MPTP correlation studies.

Introduction

Common thought on basal ganglia function has been dramatically altered in the last fifty years. As in other fields of modern science, these changes in understanding were motivated by a series of new models of the basal ganglia. These models have yielded better understanding of old knowledge, new experimental predictions, and finally have been replaced by more modern ones because of new findings and thinking. In this chapter we elaborate on one of the most prominent family of models of the basal ganglia — the action selection models (Mink, 1996). The common emphasis of the action selection models is on the role of the basal ganglia in choosing one or more actions out of a multitude of such actions presented to the basal ganglia by the cortex (Mink, 1996). The different models of this family vary in the nature of the actions selected, ranging in their definition from low-level “simple” motor actions to high-level “complex” behavioral schemes. The...
action-selection models can also be divided into two major categories according to the architecture of this selection mechanism, i.e., whether selection is performed in an intra-nucleus or inter-nucleus manner. Intra-nucleus selection is achieved using lateral (or recurrent) inhibition within the nucleus, whereas inter-nucleus selection uses feed-forward competition between the different pathways.

Models of intra-nucleus action selection stem from anatomical data showing lateral GABAergic connections between the medium spiny neurons (MSNs) of the striatum (Wickens, 1993, 1997; Bäuerle and Herz, 1998). In these models, a "winner-take-all" mechanism is implemented by way of a neural network that converges to a single winner. The dense network of inhibitory connections between MSNs is assumed to inhibit neighboring neurons, thereby maintaining the activity of only a single neuron. The selection process takes place within the layer of the striatum before transferring the chosen action via the different pathways to the output layer of the basal ganglia.

Inter-nucleus selection models vary in the ways their selection mechanisms operate and in the roles suggested for the different nuclei and pathways of the basal ganglia. Moreover, even the final outcome of the selection process, distribution of the thalamocortical networks, is performed by the basal ganglia in two basic ways, according to two families of action selection models: simultaneous or sequential activation of opposite pathways (Hoeksma et al., 1993). In simultaneous activation, the direct and indirect pathways operate on the output nuclei simultaneously, resulting in a sharpening effect and spatial focusing of the output targets. In sequential activation (also called temporal scaling), the effects of one pathway are followed (or reach) by the opposite effects of the other pathway. According to the focused selection model (Bruns and Sejnowski, 1986; Mink, 1996; Gurney et al., 2001) the basal ganglia receives input from multiple cortical sources—representing a multitude of possible actions—and enables normal motor function by releasing (or disinhibiting) a single activation pattern while inhibiting the other patterns. Using this focused selection, competing possible actions are prevented from working simultaneously and disrupting normal motor function. Thus, it is expected according to these models that the neuronal representative of the chosen action would reduce its activity at the output level (reflecting disinhibition), while all the representatives of all competing schemes increase their activity, inducing increased inhibition.

Unlike the earlier box and arrow model (Albin et al., 1989; DeLong, 1990), action selection models do not regard the different basal ganglia nuclei as uniform entities, but rather explore the internal relationships between representations within each nucleus. The models assume the existence of "units" which correspond to neuronal pools within a single BG structure. This allows new experimental predictions concerning the interactions between neighboring neurons. The "winner-take-all" nature of the system, which stems from the outcome of basal ganglia output disinhibition of the thalamocortical network, imposes strong correlations between members of opposing and complementary functional units. One would expect to find strong negative correlations between the winner and the losers, and positive correlations between the losers. Therefore, cross-correlation studies are essential to determine the different possible modes of function in the basal ganglia (Ramnani et al., 1998). Multiple electrode recordings (Baker et al., 1999) and spike sorting methods (Lowe et al., 1998) that enable discrimination between two neighboring units whose electrical activity were recorded by a single electrode are the main methods used to achieve this goal.

In this study, we aimed to investigate the predictions of action selection models regarding the responses (polarity and specificity) to various behavioral events as well as the correlation of pallidal neurons. To examine the activity of pallidal neurons during the process of action selection and performance, we recorded from monkeys performing a delayed-release Go/No-Go task. This allowed us to dissociate between trials requiring action selection, and others, identical in all other respects, that did not. Following the action selection model, and assuming that in delayed-release task an action is selected after the trigger signal, we predicted that some pallidal neurons responding with suppression of firing rate would be found following the trigger signals in the Go compared to the No-Go behavioral mode. Moreover, we expected that those cells involved in producing a single action (e.g., the winners—those reducing their firing rate) would have different responses for
different movements. Finally, strong (either negative or positive) correlations were expected in the spiking activity of pallidal neurons.

**Methods**

Two vervet monkeys (monkeys H and J: *Cercopithecus aethiops aethiops*) were used in this study. Handling of the monkeys and all procedures were in accordance with the NIH Guide for the Care and Use of Laboratory Animals (1996), and with the Hebrew University guidelines for the use and care of laboratory animals in research, supervised by the institutional Animal Care and Use Committee.

The monkeys were trained for a Go/No-Go behavioral task (Fig. 1, see for details Raz et al. (1996, 2000)). After training for the behavioral task, a recording chamber was attached to the skull with an acrylic cap to allow access to the GP under general anesthesia and aseptic conditions. During recording sessions the monkeys' heads were fixed. Eight glass-coated tungsten microelectrodes (impedance 0.3–1.2 MOhm at 1000 Hz) were individually advanced (BiP, Alpha- Omega Engineering, Nazareth, Israel) to the GP. Signals from the electrodes were amplified with a gain of 10 K and band-pass filtered with a 300–6000 Hz 4-pole Butterworth filter. Extracellular action potentials were detected and classified on-line using a sorting algorithm based on level-crossings, followed by template-matching. Spike detection signals and behavioral events were archived to a data acquisition system at 1 KHz. Cells were classified as pallidal neurons if they were found at the expected stereotaxic coordinates and by their electrophysiological characteristics.

Spike-trains were used for further analysis only if their spike waveforms were reliably separated from those of other units during the online spike sorting. Each of these spike trains was then analyzed for stability. In this analysis, the rate of each unit as a function of time was displayed graphically for the entire period of recording. Only units that were judged as steady for at least ten trials of each behavioral event (Go, No-Go behavioral modes, Cue and trigger signals for right and left directions) were considered for further analyses. Responses to behavioral events were studied by the comparison of the average firing rate after the event (100 ms after the visual cue and 1100 ms after the trigger signal) to the 50) ms before the event. The testing epochs were selected to cover the average responses to each of the events.

Cross-correlation functions (cross-correlograms) were calculated for the control and the post-MPTP periods with a 1 ms bin size for a time window of 1 s. The S1D and expected correlation were calculated for each cross-correlogram using the first 1/5 and last 1/5 of the correlogram. A correlation between a pair of neurons was considered significant if it included at least three consecutive bins that deviated from the expected correlation by at least three STDs.

**Results**

In the pre-MPTP (normal) recording 26 pallidal cells satisfied the inclusion criteria of this study. Typical responses to the inspected events are depicted in Fig. 2. In agreement with the suggested role of the basal ganglia in motor control we found that the number of responses to the Go signal (18-20/26) was greater than the number of responses to the visual cue (11/26). However, the fraction of responding pallidal neurons, as well as the ratio of positive to negative responses, was equal in the Go and No-Go modes (Fig. 3). The fraction of negative responses was significantly higher following the cue signal (47%) than following the trigger signal (22%) both in the Go and the No-Go mode (Fig. 4). This might suggest that the action selection is performed immediately after the cue signal and not following the trigger signal. We therefore tested the ability of neurons to discriminate between right and left cues (Fig. 5). Most (80%) of the cue responses were sensitive to the laterality of the cue, whereas
only 25% of the responses following the trigger signal were sensitive to the cue or movement direction (Fig. 6).

Recordings of striatal pairs in behaving monkeys (Jaeger et al., 1995) and recordings of striatal pairs in anesthetized rats (Stem et al., 1998) have shown no significant correlations on a short timescale. Similarly, our previous studies of pallidal activity revealed that less than 10% of the pallidal cross-correlograms (calculated at different behavioral epochs) displayed significant correlations in the normal monkey (Nim et al., 1995; Raz et al., 2000; Heimer et al., 2002). A study of the firing patterns of neighboring neurons shows that they have the same characteristics as tonically active neurons within the globus pallidus (GP), namely uncorrelated firing (Bar-Gad et al., 2003). Our preliminary studies of the other output nuclei of the basal ganglia — the substantia nigra pars reticulata (SNr) revealed similar non-correlated activity in this structure as well.

The development of the primate 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) model of Parkinsonism in the early 1980s (Burns et al., 1983; Wilins et al., 1989) has led to great progress in our understanding of the pathophysiology of Parkinson's disease (PD). Monkeys treated with MPTP exhibit most of the cardinal symptoms of PD, including: akinesia, bradykinesia, hypometric movements, cogwheel rigidity, flexed posture and loss of facial expression (Burns et al., 1983). Multiple-electrode studies in the GP of MPTP-treated monkeys demonstrated that the cross-correlograms become peaked and oscillatory (Raz et al., 2000). Similar findings of increased synchronization within primate brains following MPTP-treatment have been found in primary motor cortex (Goldberg et al., 2002), among striatal TANs and between the TANs and pallidal neurons (Raz et al., 1996, 2001).

**Discussion**

The action selection model of the basal ganglia suggests that the basal ganglia receives input from multiple cortical sources — representing a multitude of possible actions — and enables normal motor
function by releasing a single pattern while inhibiting the other patterns. Because of the GABAergic inhibitory projection from the output nuclei of the basal ganglia (e.g., the internal segment of the globus pallidus GPi) — the result is a “winner-lose-all” network (Berris and Sejnowski, 1996, 1998). In other words, the pallidal loser — e.g., the pallidal neurons that decrease their firing rate — disinhibits the appropriate portion of the subsequent thalamocortical network, and releases the appropriate action (causing it to win). Although most action selection models have focused on the selection of a single action, some work has been done on “soft” selection in which multiple actions may be chosen in parallel. A model that implements such an approach is the “winner-takes-all” model (Fukai and Tanaka, 1997). Similarly, the “single” action selection model can be applied to several parallel loops of the basal ganglia (Alexander et al., 1986; Middleton and Strick, 2000) and are not strictly restricted to the release of a single action.

In order to test the predictions of the action selection model of the basal ganglia, we trained a monkey to perform a delayed-release Go/No-Go task, and recorded the activity of pallidal neurons during the performance of this task. In agreement with previous studies (DeLong, 1972; DeLong et al., 1985; Turner and Anderson, 1997; Borand et al., 2000) we found that most of the responses of the basal ganglia consist of an increase in discharge rate.
rather than suppression of the spontaneous high rate (40-70 Hz) of pallidal neurons. This finding is consistent with the main prediction of the action selection models, since one can expect that there should be more actions to inhibit than the cue or the few that are being selected. In the outset of the study we assumed that actions are selected in the Go but not in the No-Go mode, and that these actions were selected following the trigger signal rather than at earlier stages of the trial. We therefore predicted that there would be more negative responses (suppression of firing rate) in the Go mode than in the No-Go mode, and that in the Go mode we would find more negative responses following the trigger signal (instructing the monkey to perform the rewarded movement) than following the cue signal (which the monkey has to neutralize for the delay period till the trigger signal). Surprisingly, we failed to see any differences in the fraction of negative responses between the Go and the No-Go mode, and we found a larger fraction of negative responses following the cue signal rather than following the trigger signal.

Our findings can still be explained in the framework of the action selection model if one assumes that the No-Go trials entail actions just as the Go trials. In these, the monkeys are required to ignore the Go signal, keep their hand on the central key, wait for future reward, etc. Furthermore, the No-Go actions are just as specific as the Go actions, in that specific actions need to be explicitly suppressed. Additionally, if they are to be reconciled with the action selection model, our findings suggest that actions are selected immediately following the cue signal, rather than directly preceding movement. Indeed, we found that the fraction of (cue and movement) direction sensitive responses was greater following the cue signal rather than following the trigger signal. However, most action selection models predict strong (either positive or negative) correlation between the spiking activity of pallidal neurons. Our multiple neurons recording in the pallidum failed to disclose such correlation (Nini et al., 1995; Raz et al., 2006; Heimer et al., 2002; Bar-Gad et al., 2003). Moreover, it is not clear why this normal asynchronous activity should transform to synchronized following the MPTP treatment and induction of Parkinson's disease.

We recently suggested the reinforcement driven dimensionality reduction model for the basal ganglia network (Bar-Gad and Bergman, 2001; Bar-Gad et al., 2004), which in our view is more compatible with the normal and pathological physiological findings mentioned above. The model is an extension of a neural-network based model for performing principal component analysis (PCA) (Foldiak, 1989; Kung and Diamantaras, 1990), which uses symmetric or asymmetric lateral interactions in the output layer. The activity of the output neurons is the weighted
sum of the feed forward inputs (weighted by the efficacy of the connections) minus the weighted sum of the lateral inputs. The learning rates are extensions of the normalized Hebbian rule (Oja, 1982) for the feed-forward weights and extensions of the Hebbian rule for the lateral network. Finally, the model adds a reinforcement factor to the supervised learning of the network. This reinforcement factor is multiplied by the input and the output of the neuron to create a multi-Hopfield learning algorithm for the cortico-striatal synapse. In a multi-Hopfield learning rule, the reinforcement signal regulates the amount of change in the weight of a synaptic connection for a given input/output pair (Reynolds and Wickens, 2002). The network displays a dynamic pattern of activity which, given inputs from a certain distribution, goes through a learning process leading to stabilization. If the input to the network (Foldiak, 1989) contains significant correlations, simulating the correlated input from the neighboring areas in the cortex (Eggermont, 1990), the output neurons initially display correlated activity since multiple output neurons encode the same aspects of the input. However, over time the initial inhibitory network causes an orthogonalization of the activity of the output neurons, leading to an uncorrelated firing pattern (Foldiak, 1989). The uncorrelated state is maintained by the pattern of efficacies of the feed-forward weights. This uncorrelated firing pattern is in agreement with the correlation studies in the striatum (Jaeger et al., 1994; Sterr et al., 1998) and the globus pallidus (Nina et al., 1995; Raz et al., 2000; Ieteren et al., 2002; Stanford, 2003; Bar-Cad et al., 2003).

In contrast with the action selection model, which emphasizes the role of the basal ganglia in action selection, the dimensionality reduction model suggests a continuous role for the basal ganglia network in the chain of sensory-response behavior leading to reward. Finally, the model suggests that the pallidal encoding is a distributive, rather than a "winner-take-all" encoding. Therefore, both increases and decreases in firing rates of pallidal neurons are critical for the transfer of the pallidal information to the thalamo-cortical networks. It is our hope that further studies of the basal ganglia networks will drive us to develop better models and better understanding of the physiology and pathophysiology of the basal ganglia and Parkinson's disease.

Abbreviations

GABA gamma amino-butyric acid
GFe globus pallidus external segment
GF globus pallidus
GFi globus pallidus internal segment
MPTP 1-methyl-4-phenyl-1,2,3,6-tetrahydro-
apyridine
PD Parkinson's disease
RDDR reinforcement driven dimensionality reduction
SN substantia nigra pars compacta
SNr substantia nigra pars reticulata
STN subthalamic nucleus
TAN tonically active neuron
VTA ventral tegmental area

Acknowledgments

We thank A. Fringold for sharing her data with us. This study was partly supported by a Center of Excellence grant administered by the ISF, the German-Israeli Binational Foundation (GIF) and the BMBF-MOS Israel-Germany collaboration in medical research.

References

Results V


Anatomical funnelling, sparse connectivity and redundancy reduction in the neural networks of the basal ganglia

Genela Morris, Alon Nevet, Hagai Bergman

Department of Physiology, the Interdisciplinary Center for Neural Computation and the Eric Roland Center for Neurodegenerative Diseases, Rambam Medical School, The Hebrew University, P.O. Box 12272, Jerusalem 91120, Israel

Abstract

The major anatomical characteristics of the main axis of the basal ganglia are: (1) Numerical reduction in the number of neurons across layers of the feed-forward network, (2) lateral inhibitory connections within the layers, and (3) neuro-modulatory effects of dopamine and acetylcholine, both on the basal ganglia neurons and on the efficacy of information transmission along the basal ganglia axis. We recorded the simultaneous activity of neurons in the output stages of the basal ganglia, as well as the activity of dopaminergic and cholinergic neurons during the performance of a probability decision-making task. We found that the functional messages of the cholinergic and dopaminergic neurons differ, and that the cholinergic message is less specific than that of the dopaminergic neurons. The output stage of the basal ganglia showed uncorrelated neuronal activity. We conclude that despite the huge numerical reduction from the cortex to the output nuclei of the basal ganglia, the activity of these nuclei represents an optimally compressed (uncorrelated) version of distinctive features of cortical information.

Keywords: Striatum, Globus pallidus, Acetylcholine, Dopamine, Parkinson’s disease

1. Introduction: main anatomical constraints

Information processing in the brain is bounded by the underlying anatomical substrate. Within the limits set by anatomy, the actual physiological parameters (e.g., firing rate, pattern and synchronization among groups of neurons) set the modus operandi for the computational processes. In this manuscript we will use the results of recent anatomical and physiological investigations of the basal ganglia to shed light on the possible computational processes and tactics used by these structures.

The detailed anatomy of the basal ganglia (see reviews in [1–3]) is beyond the scope of this manuscript. Moreover, we fear that as is often the case in discussion of anatomical detail, it may cloud the overall view of the circuit. From a broad perspective, the striatum (input stage of the basal ganglia) receives excitatory projections from most cortical areas as well as from several thalamic nuclei [4]. Subsequent direct and indirect projections link striatal neurons to the output stage of the basal ganglia, i.e., the internal segment of the globus pallidus (GPI) and the substantia nigra pars reticulata (SNr). The GABAergic inhibitory projections of GPe and SNr neurons control the activity of the excitatory thalamo-cortical connections (Fig. 1). In this report we will only consider the “main axis” of the basal ganglia: cortex–striatum and GPe/SNr. Each level of this cortex–striatum–GPe/SNr pathway is characterized by a high degree of numerical reduction in the number of neurons. The number of striatal neurons is one to two orders of magnitude smaller than that of cortical neurons projecting to the striatum [5], and an additional reduction of the same magnitude occurs from the striatum to the GPe/SNr [6–8]. Most anatomical studies concur that this axis is in fact comprised of a number of domains. However, the degree to which the different domains overlap is still under debate [10–15]. The feed-forward frame of the cortex–striatum–GPe/SNr is complicated by lateral connectivity, as well as by the action of neuro-modulatory substances. Most striatal and pallidal neurons form massive collateral GABAergic connections within their nuclei of origin [16]. Furthermore, the collateral inhibitory system of the striatum is augmented by
2. Materials and methods

2.1. Animals, behavioral task and surgical procedure

Three rhesus (Macaca mulatta) monkeys were trained to perform a self-initiated probabilistic delayed visual-saccadic task, in which the probability of receiving reinforcement for correct performance depended on the presented visual cue. During all training and recording sessions, monkeys were seated crosswise on a chair with a panel consisting of three keys in front of them. Trials were initiated when the monkey touched the central key. After a variable delay (5-25 sec in monkeys C and D, 25-40 sec in monkey Y), a visual cue appeared for a short period (150 ms in monkeys C and D, 500 ms in monkey Y) on a randomly chosen side of the screen. The monkeys were well acquainted with a set of five possible cues. Each cue was associated with a different probability of reward (0.05, 0.5, 0.75 and 1.0). The cue presentation was followed by a fixed hold period of 2 s, after which a go signal appeared (upon which the monkeys were required to press either the left or right key, according to the color of the memory cue). Correct responses were reinforced (with an interval of 960 ms), by liquid reward at the probability associated with the visual cue. All trials (correct, incorrect, rewarded and unrewarded) were followed by a variable inter-trial interval (3-6 s in monkeys C and D, 5-7 s in monkey Y).

After training, a square recording chamber with a 27 mm (inner) side was attached to the skull to allow access to the basal ganglia. The recording chamber was fixed 30° laterally in the coronal plane, with its center targeted at stereotaxic coordinates of the GPi or the SNr. The monkey's coordinates were adjusted according to MRI maps (Motor, Brain 4.2.1 Animal system; Kaas procedure; effective Hr: 80 μA and Tm: 0.5-0.8, 3 current shock 200 μs wide). All surgical and MRI procedures were performed under general and deep anesthesia. The monkeys' care and surgical procedures were in accordance with the NIH Guide for the Care and Use of Laboratory Animals (1996), and with the Harvard University guidelines for the use and care of laboratory animals in research, supervised by the institutional animal care and use committee.

2.2. Data acquisition and analysis

During recording sessions, the monkeys' heads were immobilized and eight glass-coated tungsten microelectrodes (impedance 0.3-1.3 MΩ at 1000 Hz), combined
within a cylindrical guide (0.65 mm inner diameter), were advanced separately (EPS, Alpha-Omega Engineering, Nazareth, Israel) into the recording targets (Fig. 1). The signal from the electrodes was amplified with a gain of 10 K and band-pass filtered with a 300-6000 Hz 4-pole Butterworth filter (MCP+, Alpha-Omega Engineering, Nazareth, Israel). This electrical activity was sorted and classified on-line using a template-matching algorithm (MSD, Alpha-Omega Engineering, Nazareth, Israel). The sampling rate of spikes detection pulses and behavioral events was 12 KHz (AlphaMap, Alpha-Omega Engineering, Nazareth, Israel). Eye movements were recorded with monocular infrared oculomotor (Dr. Beins, Karchrude, Germany) and sampled at 9.8 KHz. The analog output of all electrodes was continuously sampled at 24 KHz after 5 K amplification and 1-6000 Hz band-pass filtering (Figs. 2 and 4).

Only spike trains considered during real-time sorting to be emitted by a single cell were subjected to rate stability analysis. In the rate stability analysis a smoothed estimate of the instantaneous rate of a neuron as a function of time was displayed for the entire period of recording, and the longest segment of stable data was selected for further analysis. Subsequently, the cells' responses to behavioral events, autocorrelograms and pairwise cross-correlograms were calculated. The cross-correlograms were calculated only for pairs of cells recorded by different electrodes to avoid artifacts due to sorting shading [21].

3. Results

3.1. Correlation studies of the spiking activity in the output stages of the basal ganglia

Taking into account the abundant lateral inhibitory connections in the striatum one might expect strong lateral physiological interactions between striatal neurons. This prediction, however, has not been supported by previous physiological, intra-cellular studies. No evidence has been shown for functional synaptic interactions between striatal projection neurons [22]. Recent in vitro studies, using spikes triggered average techniques, revealed a small number of weak inhibitory connections between striatal projection neurons, and those found were all uni-directional [23].

It is extremely difficult to perform intra-cellular studies in the awake behaving animal. Therefore, we used multiple extra-cellular electrodes to record the simultaneous spiking activity of several neurons in the output stages of the basal ganglia. Our working assumption is that functional lateral inhibitory connections should result in troughs in the cross-correlation functions. Functional connectivity in the form of striatal or pallidal functional assemblies, created by shared inputs, should be evidenced as positive peaks in the cross-correlation functions of neurons in the respective nuclei. Previous studies have failed to reveal correlations between the spiking activity of simultaneously recorded pallidal neurons [24,25]. It should be noted, however, that the lateral connectivity, as well as the number of interneurons and the direct dopaminergic effects in the SNr show distinct differences to those observed in the GPi [26,27]. We therefore set out to study the correlation between simultaneously recorded SNr neurons (Figs. 1 and 2). Our preliminary analysis reveals that as with GPi, the pairwise correlation of SNr neurons is relatively flat (Fig. 3). Such correlation functions suggest that the amount of functional connectivity within the SNr is minimal, as is the degree of convergence of common input to SNr neurons.
3.3. Teaching signals in the basal ganglia

It has consistently been shown that the striatum shows the densest tracing in the central nervous system for dopaminergic [1] and cholinergic markers [28]. While most of the brain dopamine is generated by midbrain dopaminergic neurons and merely projected to the striatum, striatal acetylcholine is probably generated by the striatal cholinergic interneurons. There are several studies that indicate that the tonically active neurons (TANs) of the striatum are the cholinergic interneurons of the striatum [29-31].

The central role of dopamine and acetylcholine in control of motivation and learning has been known for many years [72]. Recent studies [33,34] revealed that the dopaminergic signal is less characterized as related to the discrepancy between the animal's predictions and reality [35]. Thus, dopamine neurons respond only to the first cue in a trial that predicts reward. However, they do not respond to aversive stimuli [36].

Initial physiological studies of TANs revealed similar properties to those of the dopaminergic neurons, with inverse polarity of responses [37]. However, several recent studies indicate that the functional message of the cholinergic system may be different from that of the dopaminergic system. Thus, the TANs show robust responses to aversive stimuli [38] and respond to more than one event in a trial [39].

In our recordings, most TANs exhibited a stereotypical response to the visual cue (a pause in firing, followed by a brief elevation of the firing rate). The same cells also displayed a similar response to the reward (Figs. 4 and 5). However, the TANs' response was not significantly modified by the different cues (right or left side and different probability for future reward). Thus, unlike dopaminergic neurons that code for the difference between the animal's prediction and reality [39], the TANs provide a more general message, probably indicating, or alternatively instructing the occurrence of an attention shift to salient events.

4. Discussion

4.1. Reinforcement learning models of the basal ganglia

The "actor-critic" architecture is a standard learning system [40], in which there is an "actor", acting in a certain environment, and a "critic", providing reinforcing signals which are used by the "actor" in order to maximize the weighted sum of all future reinforcement values.

The apparent activity in the basal ganglia of the reward system resembles that of the "critic" in reinforcement temporal delay learning models [41]. Therefore, in computational models of the basal ganglia [33,42-44], the cortex striatum GPe/SNr acts is functionally modeled as the "actor", while the dopaminergic (and cholinergic) neurons are processed as the "critic" or the provider of an error signal in a learning network.

Actor critic network models postulate that the teaching signal will modulate synaptic transmission in the actor. Indeed, it has been shown that plastic changes in the morphology of striatal synapses occur after dopamine depletion [45]. Physiological studies show that the dopaminergic [46,47] and the cholinergic [48] signals modulate the cortical input to striatal projection neurons. Moreover, as predicted by reinforcement learning models, striatal and pallidal neurons significantly change their discharge as a function of the prediction of future rewards [49,50]. Furthermore, their discharge patterns vary significantly during different phases of learning [51].
Fig. 7. Multi-electrode recordings of TANs in the striatum. An example of 3 s of the simultaneous output of five electrodes positioned in the striatum of a normal behaving monkey. The "electrode output was sampled at 20 kHz and digitally band-pass filtered at 500–6000 Hz.

The balance between the dopaminergic and the cholinergic systems in the basal ganglia has been studied extensively since the discovery of the beneficial therapeutic effects of anti-cholinergic drugs for parkinsonian patients [53]. Anatomical studies of the striatum revealed direct contacts of dopamine terminals and striatal cholinergic neurons [54]. Furthermore, numerous neurochemical studies demonstrated that dopamine application inhibits acetylcholine release within the striatum. Intriguingly, it has recently been demonstrated that acetylcholine has opposing effects on the release of dopamine in the striatum [55]. Experimental procedures that abolish dopamine input to the striatum (Haloperidol, MPTP [56,57]; local application of D2 antagonists [58]) abolish the TAN's response. However, re-establishment of the TAN activity by treatment with non-specific (post-synaptic) dopamine replacement therapy [56] indicates that the dopaminergic system enables (but does not drive) the TANs to express their stereotypical responses.

Our results show that the functional message of the cholinergic system is not identical to that of the dopaminergic system. Unlike the dopaminergic activity, which is proportional to the difference between prediction and reality, the TAN message does not depend on the probability of future reward. TANs show a similar response for both unpredicted positive and negative (disadvantageous) events (data not shown). Further studies are needed to establish whether the dopaminergic and cholinergic projections converge on the same population of striatal neurons or rather on separate ones, and to establish their specific role in controlling the basal ganglia rates in the cortico-striatal synapses.

4.2. Redundancy reduction neuronal networks

The complex anatomical and physiological setting of the "active" elements in the basal ganglia (i.e., the main axis) can be combined within a computational model of local competitive learning rules [59], controlled by reinforcement or error signals [60,62]. The models assume that the basal ganglia perform efficient dimensionality (redundancy) reduction [59,60,63-65] and decorrelation of the large information space spanned by the activity of the cortical-striatal neurons. It has been proven that neural networks can perform such efficient coding using local cellular competitive learning rules [63]. Most sensory systems that have been shown to perform dimensionality reduction [66], do so solely in relation to the input statistics. Due to the added value of the "active" elements in the basal ganglia, dimensionality reduction performed in this structure is a function not only of the statistical properties of the cortical (input) patterns, but also of their behavioral significance. This is achieved by a triple striatal synapse, in which the teaching (dopaminergic or cholinergic) signal controls the feed-forward cortico-striatal (and striatal pallidal) Hebbian learning. Thus, desynchronization of basal ganglia activity is achieved by a distinct process rather than by fixed cortico-striatal GPi/SNr connectivity. More
For Doctoral Thesis, Genela Morris

Results V

![Diagram](image.png)

Fig. 5: Example of RAS responses to visual stimuli and reward probability. Each row represents one of the conditions depicted in Fig. 2, and columns correspond to different stimulus conditions, with a bar indicating the stimulus and the probability of reward.


discriminability, this network enables a discriminative reinforcement learning model, performing better for rewarded stimuli than for non-rewarded events.

4.3 Spurious connectivity and information in the basal ganglia circuit

Alternatively to the dimensionality reduction models, the apparent lack of temporal correlation between the neuronal activity within the GPI and SNr could be accounted for by spurious cortico-striatal and striato-pallidal connectivity. Recent studies indicate that the anatomy of the cortical-striatal pathway is highly and discontinuous, and that individual cortical terminations may be separated by extended distances.

Quantitative analysis of single neuron tracks reveals that the degree to which cortical input is shunted by nearby striatal neurons is low (8,6). Wilson and colleagues have shown that the density of striatal projection neurons (which is, incidentally, of an equal size to that of the local striatal internuncial connections) is high (8,66,89). This highdensity makes the striatal network a logarithmically dense network with an average basal ganglia circuit.
probability of two pallidal cells to share common striatal inputs is also very low [71].

4.4. Redundancy reduction vs. sparse connectivity

The number of cortico-striatal neurons exceeds the number of striatal neurons by a factor of 10 [5]. It was therefore concluded that unless the cortico-striatal output is overwhelmingly redundant, the striatum cannot compress the full cortical information. However, dimensionality (redundancy) reduction networks do not always perform compression without information loss. In many cases, especially in those where the number of elements in the output structure is significantly smaller than the number of elements in the input layer, the compression process will result in information loss. However, the network will keep the "most important" parts of the input patterns.

Regardless of the underlying reason, it is clear that the cortico-striatal physiological message is not a simple read-out of the cortical state [72,73]. Multiple lines of evidence suggest that cortico-striatal and cortico-spinal (cortico-pontine) neurons belong to distinct populations, and that the signal transmitted by cortico-striatal neurons is distinct from that sent to the spinal cord or the brainstem. It seems that the firing of cortico-striatal neurons is more selective than that of cortico-spinal neurons. Although there is no evidence of a correlation between cortico-striatal neurons, one may conclude that only sparse and selective cortical information is transmitted to the striatum, and therefore this information can be compressed in the much smaller number of striatal, and subsequently, pallidal neurons, only if some of it is lost.

Even more important is the finding of sparse connectivity within the basal ganglia. Each single striatal neuron receives 5000 cortico-striatal synapses. A single striatal cell is therefore exposed only to less than 9.6% of the cortical information. Standard competitive networks have "all to all" connectivity, i.e., all neurons in the output layer receive projections from all input neurons (Fig. 8(A)). Such all-to-all connectivity enables the system to adapt to changes in the input pattern, and to create optimal (with minimal loss of information) representation of the input patterns. However, more realistic redundancy reduction models of the basal ganglia must assume that each striatal neuron receives information from a limited number of cortical neurons (Fig. 8(B) and (C)). This subset of connections could either constitute one of a very few segregated cortico-striatal pathways (Fig. 8(D)), or of partially overlapping circuits (Fig. 8(C)). Finally, one does not have to assume a symmetric network. Several directionally (Fig. 8(D)) may enable the sharing of information to the basal ganglia in "ascending" order [74,75], e.g., from limbic to cognitive to motor domains.

Fig. 8. Distribution of the feed-forward connectivity to the basal ganglia. The figure shows the possible configuration for the feed-forward connectivity, (A) standard connectivity; (B) fully segregated connectivity; (C) mixed convergence/divergence network, (D) symmetric (from left to right) convergence/divergence network.

5. Conclusions

The physiological evidence, indicating lack of correlation between neurons in the output nuclei of the basal ganglia, and the massive reduction in the number of neurons from the input to the output structures, suggest that the basal ganglia output is a compressed form of certain aspects of cortical activity. However, this compression probably involves information loss. The resulting signals of the basal ganglia, delivered by cholinergic and dopaminergic neurons may enable the system to keep the most important aspects of the information. We showed that these two "critical" are different in their sensitivity. What is the significance of each of these signals in the information processing in the basal ganglia, and how the cortex uses the compressed output of this system, are yet to be answered.

Acknowledgements

This center of excellence (800/600) was supported by the Israel Science Foundation. This research was also supported in part by the United States-Israel Binational Science Foundation, the German-Israel Binational Foundation (GIF) and the EMDE Israel-Germany collaboration in medical research. G.M. and A.N. have made equal contributions to this work.

References

Doctoral Thesis, Genela Morris

Results V


Discussion

The results presented in this thesis point to a number of novel findings regarding physiology, behaviour, computational models and the interplay between them. I will first outline the main findings, and then provide a short discussion of some of their implications. I will first divide the findings according to the different cell-types that were analyzed, and in the discussion I will elaborate on some of the findings that leave interesting open issues, as well as try to provide a more unified framework.

1. Results regarding dopamine neurons:

   a. Dopamine neurons fire in a temporally independent manner.

   b. The activity of dopamine neurons during instructed choice instrumental conditioning reflect accurately the positive TD error signal used as a reinforcement signal in the reinforcement learning algorithm of TD learning.

   c. Behaviourally, I have shown that monkeys engaged primarily in instructed choice instrumental conditioning, choose a probability matching strategy when faced with a decision context.

   d. The behavioural probability matching strategy is accurately predicted by the average activity of dopamine neurons, and not by the relative average rewards.

   e. Within the decision context, the activity of dopamine neurons upon presentation of the conditioning stimulus reflects the value of the action that will be taken in the future.

2. Results regarding tonically active neurons (TANs):

   a. TANs fire as a largely synchronous population.
b. TANs respond to the same behaviourally significant events as do dopamine neurons with stereotypical responses comprised of pauses in firing flanked by increases in firing probability.

c. The responses of TANs are time-locked to those of the dopamine neurons.

d. The responses of TANs to different conditioning stimuli is invariable with the value of the stimuli, and the responses to reward and its omission are invariable with reward predictability.

3. Results regarding globus pallidus (GP) neurons:

a. Globus pallidus neurons fire independently in the healthy monkeys

b. The activity of these neurons during instructed choice tasks conforms to an action selection mechanism, but not to action execution.

**Dopamine neurons reflect an accurate TD-error signal in the positive domain**

When the monkeys performed instructed-choice instrumental conditioning tasks, the activity of dopamine neurons displayed an increase in firing during the time of CS presentation which was directly proportional to the value of the stimulus. Upon subsequent reward delivery the neurons responded with a phasic elevation of firing which was inversely proportional to the stimulus value (indicating accuracy of prediction). When expected reward failed to occur, the neurons responded with a reduction in firing. However, this reduction was relatively uniform, regardless of the level of reward prediction. This response, which was seen both on the level of single units and on the population level, corroborated the detailed predictions of the model associating dopamine neurons with the TD-error, the hypothesized reinforcement signal in the TD learning reinforcement learning algorithm. Similar results have been recently reported in monkeys who were classically conditioned to associate cues with various reward probabilities {Fiorillo, 2003 10773 /id} (Satoh et al., 2003).
The first point worth noting, however, is that the alignment of the dopamine signal with the TD-error is truncated. While for the positive errors ('good surprises') the dopamine neurons give an accurately scaled TD-error signal, this signal for negative errors ('disappointment') transforms to a digital signal, informing the system of the disappointment, but not of its extent. This deviation from the model is probably a consequence of the low-frequency spontaneous discharge of the dopamine neurons. Negative TD-errors require further reduction in firing, which is naturally truncated by 0, and therefore efficient rate coding of the negative domain is highly unlikely. Notably, in experimental conditions where the monkeys were explicitly informed that a negative prediction error has occurred, DA neurons seemed to report negative errors far less efficiently than positive errors, e.g., smaller fraction of neurons with pure depression of discharge and smaller magnitude or gain of the negative responses (Tobler et al., 2003; Satoh et al., 2003). Nevertheless, we wish to believe the striking resemblance of the dopamine signal to the positive TD-error is more than just an epiphenomenon, and that it reflects a real computation performed in the striatum. We are therefore left with two options of signalling the negative errors. One, that negative errors are signalled by a different teacher (e.g., serotonin in the striatum (Daw et al., 2002), or even in a different structure, such as the amygdala). A possible candidate for the negative teacher could be serotonin (Daw et al., 2002), which is abundant in the striatum (Van Bockstaele et al., 1993; Prado-Alcala et al., 2003). An alternative option would be abandonment of the rate coding for that of duration (Parush et al., 2006).
For this signal to be implemented in learning, learning itself should be scaled in some manner by this differential dopamine activity. Abundance of evidence has demonstrated that dopamine is essential to plasticity in the cortico-striatal synapses (Reynolds et al., 2001; Calabresi et al., 1998; Calabresi et al., 1999a; Calabresi et al., 2000; Centonze et al., 1999a; Centonze et al., 1999b; Centonze et al., 2001; Centonze et al., 2003). However, for our graded dopamine signal to hold meaning for learning, the effect on striatal plasticity must also scale with the amount of dopamine released. To the best of my knowledge, such an effect has not been systematically investigated. Nevertheless, regardless of the results of this hypothetical experiment, one can still envisage a graded effect on learning is achieved through the stochastic nature of binary processes. It may be that increased levels of dopamine enhance the probability of each synapse to undergo a long term effect, thereby increasing the overall level of potentiation (or depression) in the appropriate circuit.

That a physiological neural network should follow so closely the behaviour of a computational component in a class of algorithms developed purely for purposes of computer science is remarkable, and gives rise to perhaps more pretentious examination of specific computational models and algorithms for their ability to describe this structure. The results from this instructed-choice experiment are in line with the actor/critic model of the basal ganglia circuitry. This model is based on a 'critic' with no knowledge of the actions, who can track the average value of situations (called 'states'); these values can be used as rewards to train an 'actor' who is responsible for knowing the policy and making choices. This notion was particularly attractive since it falls naturally within the anatomy of the basal ganglia (Barto, 1995; Joel et al., 2002; O'Doherty et al., 2004). However, the use of an actor/critic type network by the basal ganglia has a number of predictions that can only be tested in a decision situation, to which we shall turn next.

**Monkeys use a probability-matching decision strategy**

The behavioural results in this thesis indicate that similar to human adults (Wolford et al., 2000; Vulkan, 2000), and in contrast to pigeons and rodents (Bailey
inexperienced gambler monkeys decide between two alternatives in an approximate probability matching strategy. As hinted in the introduction, this seemingly irrational behaviour may be accounted for as an attempt in solving the exploration/exploitation problem in reinforcement learning (Sutton and Barto, 1998). Dealing with the exploration/exploitation balance problem implies sub-optimal behavioural strategy. This essentially means that a reinforcement learning network must include a source of randomness to be able to explore the set of possible outputs until an optimal one is found. Randomness can be implemented in many ways. A simple way would be to design a policy in which a pre-determined fraction of the trials ($\varepsilon$) is designated to exploration, while the other trials the 'greedy' action (i.e., which will provide the highest expected reward) is taken. These policies are called $\varepsilon$-greedy (Sutton and Barto, 1998; Lee, 2006). Alternatively, the exploration fraction can be a more complicated function of the expected loss from such exploration. One such function could be the following soft-max function:

$$P_a = \frac{e^{Q_a / \tau}}{\sum_b e^{Q_b / \tau}}$$

where $Q_a$ is the value of the action $a$, and $b=1..N_b$ represents all other actions. $\tau$ is a temperature parameter. When the temperature approaches infinity, all actions are equi-probable (Sutton and Barto, 1998). If it is set to zero, a maximization policy is achieved. Intermediate temperatures determine the slope of this logistic function, and hence the range of reward probabilities in which the function is approximately linear and thus approximates probability-matching. Our results resemble an intermediate temperature value. However, the effect of temperature on the shape of the choice curve depends on the range of the values, and therefore extracting an absolute number for this value would hold little meaning.

**The decision strategy is shaped by dopamine neurons**

The classical view of reinforcement learning describes behaviour in light of reward history. Testing whether dopamine mediated plasticity is the substrate by
which reward shapes behaviour faces an absurd methodological complication: the activity of dopamine neurons is so accurately correlated with reward, that it becomes extremely difficult to dissociate between the two and find which factor is indeed the mediator of behaviour, and which is merely an epi-correlate. Still, statistical tools have enabled me to show that it is through the activity of dopamine neurons that the probability matching strategy is achieved.

Although an abundance of previous works demonstrated a connection between dopamine and behaviour reinforcement (for review see (Wise, 2004)), their vast majority was oriented towards the topic of drug dependence. Therefore, they mainly focused on paradigms such as self-stimulation and self-administration. In these works it was shown that behaviour that leads to the increase in meso-limbic dopamine activity is reinforced. This reinforcement is dependent on intact dopaminergic transmission. Other works demonstrated through lesions that dopamine is indeed necessary for learning reward-oriented behaviours. It was also shown that such behaviour is paralleled with dopamine-dependent long term plasticity (Reynolds et al., 2001). However, I believe that in this work I provided the first demonstration of the detailed correlation between specific dopamine activity and a specific behavioural change.
Dopamine neurons reflect future decisions

As for the actual online decision process, my results indicate that dopamine neurons are not likely to serve a role. This seemingly negative conclusion is paradoxically based on a positive finding; namely on the encoding of the future action by the dopamine neurons' response to presentation of the CS. In section III of the Results chapter, I showed that these responses are modulated by the value of the action that will be taken at the trial completion. This result imposes a clear bound on the position of dopamine neurons in the hierarchy of decision-making. Functionally speaking, for a structure to be considered a ‘decision maker’ it must do more than reach the decision independently. It should also serve as the source for the outflow of the decision to the appropriate executers. Therefore, the verdict on whether or not the dopamine neurons are the decision makers lies on their ability to convey decision information onto the action circuits.
A naïve solution would lie in a hard-wired mapping of dopamine neurons onto subsets of cortico-striatal connections representing different actions. In this case, the dopamine neurons would have to provide differential signals to each target population. In fact, neither prediction holds: anatomically, the arborisation of dopamine neurons in the striatum support an information divergence-convergence pattern. Specifically, the broad spatial spread and the enormous number of release sites (~5 × 10^5) of each dopamine axonal tree, additionally imposes extremely high convergence on single striatal projection neurons (Wickens and Arbuthnott, 2005). Volume transmission of dopamine in the striatum also enforces population averaging of the dopamine signal on the level of the single target striatal neuron. Finally, the lack of temporal correlations of the spiking activity of dopamine neurons (Morris et al., 2004) provides an optimal substrate for such averaging to yield accurate estimation of the transmitted signal (Zohary et al., 1994). The other prediction, namely that the dopamine neurons provide differential TD signals to different circuits is clearly invalidated by my results. Nevertheless, one can still think of a way that the dopamine neurons will aid the decision that is in line with the results presented in this theses. A subset of computational models of decision making has adopted this view (McClure et al., 2003). According to these, when an action is considered, the TD-error signal provided by the dopamine neurons is used to determine the probability with which the action will be taken (e.g., \( P_a = \frac{1}{e^{-\delta(r)-b}} \)), where \( \delta \) denotes the TD error and \( b \) is a bias term.

This view is consistent with my results which have shown correlation between future decision and the dopamine TD signal. However, this strategy would fail in a multiple–choice task. For a system to utilize this decision-making strategy, any problem should be reduced to a single yes/no question. One possibility would be to determine in this fashion the probability of choosing the most profitable action, or the probability of exploitation. If this is indeed the case, a straight-forward prediction would be that in a task involving more than two alternatives, all the 'exploration' alternatives will be chosen at random.
Otherwise, it seems much more parsimonious to undertake the somewhat less glamorous interpretation of the results, according to which the decision is taken in a different structure and the dopamine neurons are informed of it (very rapidly) to provide the cortico-striatal circuit with a decision-specific TD error signal.

This result also has serious implication for the computational modelling of physiologically feasible decision making in a temporal difference reinforcement learning framework. The reinforcement learning models differ in which value is actually the target of learning. One type of algorithm is based on 'state' learning. To translate to our decision setting, the apparent value of a trial at during the presentation of the CS is an average value of the two options (possible weighted according to the behavioural policy). This is the type of solution that is taken in the heavily modeled actor-critic algorithm. This suggests that a 'critic' with no knowledge of the policy can track the average value of states; these values can, separately, be used as rewards to train an 'actor' that just makes choices. However, this is clearly not the case according to the presented results. Instead, they demonstrate that the dopamine signals separately reports the value of either action at a state, or the Q-values. Here, again the algorithms divide to two groups, one learning 'on policy' – that is, learning while doing, and the other 'off policy' group that separates learning from the actual behaviour. These results seem to favour the first, simpler approach, which includes algorithms like SARSA (Rummery and Niranjan, 1994), because of the tight coupling between dopamine responses and trial-by-trial behaviour. However, some caution is probably warranted here since our results were based on extensive averaging.
Striatal TANs report the occurrence of behaviourally significant events, Or: Dopamine-acetylcholine balance, revisited

In chapter I of the results section I compared between the responses of dopamine neurons and of striatal (probably) acetylcholine carrying TANs to various events within the framework of the TD learning concept. This comparison was driven by the 'acetylcholine-dopamine balance' theory of striatal function and dysfunction (Barbeau, 1962; Nisenbaum and Kitai, 1995). Accordingly, given the opposite phenomenology of the responses of the two cell types, a more detailed comparison was expected to yield mirror images. Such a relationship was not found, and therefore a search for a meaning of the two messages was called for.

When attempting to decipher the message of a neuronal population two questions must be answered: when is the message given, and what does it say. First, I would like to emphasize the similarity between the dopamine and TAN populations: the when. In fact, I showed that the responses of these two neuronal types exactly coincide even on a single-neuron, single-trial level. Nevertheless, the content of their message is qualitatively different. This different is most apparent in the responses to the extreme events of reward Vs. reward omission. While the dopamine signal reverses in polarity reflecting the hypothesized changed from a positive to a negative prediction error, the response of the TANs is always similar – a robust pause in firing surrounded by less pronounced increases in firing probability. This response pattern remained consistent when the reward probabilities (and hence the prediction levels, or the expected value) were varied, indicating that in contrast to the dopamine neurons, the TANs do not report a TD error signal. This result contradicts previous thought concerning these neurons, according which (although seldom explicitly stated) the TANs were a mirror image of the dopamine neurons (Apicella et al., 1991; Apicella et al., 1998; Watanabe and Kimura, 1998; Aosaki et al., 1995). The results presented in this theses rescue these neurons from the apparent redundancy inherent to this notion.
However, it would seem at first glance that their responses are even more expendable, since they provide only binary information, which can also be extracted from the dopamine responses. What is then unique to the message given by the TANs? I believe that a clue to this question lies in temporal analysis of the TANs. Such an analysis reveals, that these neurons are strongly synchronized (Raz et al., 1996; Raz et al., 2001). This result was reproduced in my studies as well. This means that there a population, which is spread throughout the striatum responding synchronously at the time of the events that are important for reinforcement learning. These responses correspond both in time and in space (Zhou et al., 2003) to those of the dopamine neurons. I would suggest that the role of the TANs in this context is reminiscent of the beginning of this chapter: to process a message one must know 'when' to be able to extract the 'what'. I propose that the dopamine neurons give an accurate message in terms of the TD learning algorithm, but given their high variability and poor spatial and temporal precision (Cragg et al., 2000; Roitman et al., 2004; Venton et al., 2003) the time frame for the actual reinforcement of the cortico-striatal synapses is highly inaccurate. This is in sharp contrast to striatal ACh, which is rapidly degraded by the extremely dense AChE (Zhou et al., 2003). These responses may have the capacity to provide the much needed time frame for the reinforcement learning to occur.

In the computational context, I believe that the benefit of a 'when' signal is apparent. But, how would physiology implement it? Recall that reinforcement learning corresponds to LTP and LTD in the cortico-striatal pathway. It was shown that induction of plasticity in this pathway is mediated by activation of dopamine D1/D5 receptors (Reynolds et al., 2001; Kerr and Wickens, 2001). Activation of M2 muscarinic acetylcholine receptors (that are co-localized with D1/D5 receptors (Zhou et al., 2003)) reduces LTP at cortico-striatal synapses (Calabresi et al., 1998). Release from this block by the pause in the TAN pause response can serve as the window for dopamine-dependent plasticity. Furthermore, recent work has emphasized the selective influence of presynaptic nicotinic acetylcholine receptors residing on dopaminergic neurons on the amount of dopamine release (Cragg, 2006). Finally, ACh reduces the sensitivity of striatal projection neurons to their cortical inputs by fixing their up/down state (Akins et al., 1990). The TAN pause enhances this sensitivity in the critical periods in which this entire circuit is susceptible to long-term
change. This may enable striatal neurons to adjust their state according to the cortical inputs; thereby ensuring that the DA teacher reinforces the correct state of the network. This scheme is obviously an over-simplification of what is probably a very complex interaction of the two neuronal populations. The translation of 'dopamine acetylcholine balance' into physiological terms has been approached on many levels. It has been suggested, for instance, that the TANs' response may enhance the release of dopamine in each given response (Cragg, 2006), that the response of the TANs is shaped by via long term plasticity from the responses of the dopamine neurons (Reynolds et al., 2004). Given the many different types of ACh receptors and of DA receptors in the striatum, it would be conceited to presume that any of the above listed effects underlies reinforcement learning or other aspects of the interaction as exemplified in the pathology of Parkinson's disease. However, regardless of implementation, I believe that the for the readout apparatus the crucial question is the type of information encoded by the activity of the different neurons, and to this I hope that I have opened a window.
GP neurons may participate in action selection but not in action execution

Chapter IV of the results section provides an indirect examination of the action-selection model of the basal ganglia [Mink, 1996 6061 /id] [Hikosaka, 1998 9964 /id]. According to this model competing actions are continuously suppressed by the ongoing inhibition provided by the output nuclei of the basal ganglia. When a particular action is chosen it is disinhibited by a transient decrease in the activity of the inhibitory tonically firing neurons in these nuclei. Different versions of the action-selection model of the basal ganglia have become increasingly popular, as they fit well with both physiological and pathological knowledge of the function of these nuclei. Nevertheless, the term 'action selection' is somewhat misleading, since these models discuss, in fact the execution of the selected movement rather than selection per se. In this chapter, I examined the response selectivity of GPi neurons to different types of action in two conditions – GO, which involved action, and NO-GO, which did not involve action. The study showed that GPi neurons could be involved in coding the correct action soon after the decision, rather than at the time of execution. It seems, therefore, that assigning these neurons with the 'release of action' role would be artificial, since action is not released at the time of the dis-inhibition.

Indirectly, this study complements the one reported in chapter III of the Results section in placing the final nail in the coffin of the actor/critic models of the basal ganglia. These models assume that the main basal ganglia axis, ending at the output nuclei and their relay to the motor areas of the thalamus and cortex are the holders and executers of the policy. The results presented in results chapter IV cast doubt on the 'actor' abilities of a crucial junction in this axis – the GPi. This is also very much in line with classic electrophysiological studies comparing the timing of movement related activation of various motor structures [Crutcher, 1990 3021 /id]. It seems that the part of the basal ganglia formerly known as the 'actor' can by no means be responsible for initiating the cascade of action execution.
In this thesis I focused mainly on the neuromodulators in the basal ganglia. I showed that the neurons transmitting dopamine and acetylcholine act in concert to ensure optimal tuning of the efficacy of cortico-striatal synapses in a manner that is consistent with formal theory of reinforcement learning. I further showed that this learning is indeed subsequently implemented in the animals' behaviour. However, a comprehensive view of what it is exactly that these cortico-striatal connections and the subsequent basal ganglia circuitry achieve within the greater scheme of the (no longer motor) circuitry of action production remains illusive. I believe that it is now clear that straightforward action selection and execution cannot be the case. In chapter V of the results I show one possibility for the integration of the neurophysiological activity of dopamine and acetylcholine with the known physiological, pathophysiological and anatomical properties of the remaining basal ganglia nuclei. This representation is admittedly flawed in some aspects, and in any case too simplistic to provide useful insight. Nevertheless, it is this type of model of the basal ganglia that is now pressing to establish, particularly due to the obvious benefits of understanding a structure that is so clearly involved in disease.
Reference List


