A computational account of latency impairments in problem solving by Parkinson's patients

Patrick Simen (psimen@math.princeton.edu)
Applied and Computational Mathematics / Center for the Study of Brain, Mind and Behavior; Princeton University

Thad Polk (tpolk@umich.edu)
Department of Psychology; University of Michigan, Ann Arbor

Rick Lewis (rickl@umich.edu)
Department of Psychology; University of Michigan, Ann Arbor

Eric Freedman (freedman@umich.edu)
Department of Psychology; University of Michigan, Flint

Abstract

Cognitive slowing is a feature of problem solving performance by Parkinson's patients. Here we present a computational model of the Tower of London problem solving task that provides a straightforward explanation of this latency impairment. The model is a neural network consisting of a set of idealized neurons whose activation levels are continuously varying quantities ranging between 0 and 1, organized hierarchically into a collection of structured, columnar assemblies. The model differs from many neural network models in that it intentionally approximates the discrete stages of processing and rule-based computation of production systems. It differs from many symbolic models in that it does not assume an instantaneous transition between different stages of processing and does not rely on a central, system-wide clock signal. As a result, the model faces timing problems that can prevent it from approximating an idealized discrete-time system closely enough to perform symbolic computation. The columnar assembly provides a solution to these problems by imposing a direction on the flow of activation between units and by controlling the rate of that flow. Columns are proposed to play the role of distributed timers implemented in cortical columns and frontostriatal loop circuits in the brain. Dopamine depletion in Parkinson's is proposed to reduce the rate of information transmission through a frontostriatal timer circuit critical for the generation of subgoal representations in prefrontal cortex. The model fits latency impairments in problem-solving by Parkinson's patients relative to controls and predicts that only problems that require the generation of subgoals will produce a significant latency impairment in Parkinson's patients.

Contribution of prefrontal cortex and basal ganglia to problem solving

Many researchers have noted the similarity between cognitive impairments in patients with Parkinson's disease (PD) and patients with lesions of the prefrontal cortex (PFC). Both groups show impairments related to problem solving, planning and set-shifting (Owen *et al.*, 1992; Owen *et al.*, 1995; Robbins & Rogers, 2000). Both groups are impaired on the Tower of London problem solving task depicted in Fig. 1 (which is the task that we model). Prefrontal patients have difficulty achieving an optimal solution to the puzzle (a solution trajectory that involves a minimal number of moves), but only when the required number of moves to solution exceeds three (Owen *et al.*, 1990, Shallice, 1982). The same pattern of impairments has been seen in severe cases of PD (Owen *et al.*, 1995) (although another experiment showed no such accuracy impairment (Owen *et al.*, 1992)). Here we propose a computational account of the function of brain structures affected in both conditions that may explain the similarities and differences of these cognitive deficits.

PD involves the death of dopamine-producing neurons in the substantia nigra that project to the striatum, and it is this anatomical fact that suggests a connection between the effects of PD and prefrontal lesions. Prefrontal areas of cortex are connected back to themselves by multi-synaptic 'frontostriatal loops' whose first segments consist of excitatory axonal projections from deep cortical layers to the striatum in the basal ganglia. These striatal neurons send inhibitory projections to neurons in the globus pallidus and subthalamic nucleus (also considered structures of the basal ganglia). These in turn inhibit neurons in the thalamus that send excitatory projections back to the areas of cortex near the starting points of the loops. At each stage, a notable degree of topographical organization is seen, such that the loops can be considered to be, to some degree, segregated and parallel (Alexander, DeLong & Strick, 1986). Functionally, signals from cortex to the basal ganglia have the effect of disinhibiting the thalamus and exciting cortex in a form of positive feedback.

In the striatum, dopamine appears to potentiate the excitatory effect of cortical glutamate received by medium spiny neurons, as well as to modulate synaptic plasticity (Alexander *et al.*, 1986). A natural hypothesis, therefore, is that the disruption of frontostriatal loops (which are one means by which PFC communicates with itself) in PD mimics the effects of prefrontal lesions, and that this is why the two patient groups exhibit similar cognitive deficits. However, while similarities exist between prefrontal and Parkinsonian cognitive impairments, there are salient differences. In all levels of PD severity, impairment in problem solving latency (the time to begin problem solving once a problem has been presented) is seen (see Fig. 2). Interestingly, this impairment again appears only on problems requiring more than three moves. But a similar latency effect is not seen in prefrontal patients (Owen *et al.*, 1992; Owen *et al.*, 1995).

We have previously argued that the primary role of dorsolateral PFC in problem solving is to represent subgoals that guide the selection of actions (Polk *et al.*,

2002), thereby providing a potential explanation of suboptimal problem solving in prefrontals as resulting from a failure to select actions consistent with subgoals. In this paper, we argue that latency impairments observed in PD arise as a result of slowed generation of subgoals after action selection impasses occur.

Tower of London task

The Tower of London (TOL) task, is shown in Fig. 1. This task has been used extensively to assess planning impairments and is thought to depend crucially on goal management (Shallice, 1982). It is a variant of the Tower of Hanoi problem and involves moving colored balls on pegs from an initial configuration until they match a goal configuration. Unlike the Tower of Hanoi problem, there are no constraints specifying which balls can be placed on which others, but the pegs differ in how many balls they can hold at one time (the first peg can hold one ball, the second peg can hold two, and the third peg can hold three). There is typically one red, one green and one blue ball. Participants are often asked to try to figure out how to achieve the goal in the minimum number of moves and are sometimes asked to plan out the entire sequence of moves before they begin (Owen *et al.*, 1990; Shallice, 1982).

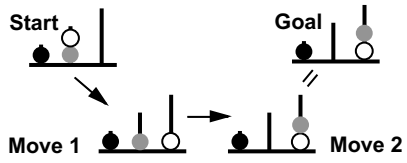

Figure 1: The Tower of London task.

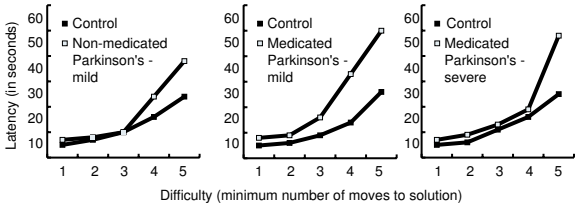

Figure 2: PD patients display latency impairments (slower planning prior to first move) in the Tower of London task. Increased latency relative to control subjects is shown by patients at two different stages of the disease, with and without medication in the early stage.

Components of a neural cognitive architecture

For the most part, the model we present works by implementing in neural networks some of the computational assumptions of symbolic production systems. We have argued previously that there is a natural mapping from such symbolic systems onto the brain (Polk *et al.*, 2002). Production systems typically assume a production matching cycle that occurs at regular intervals governed by a central clock, and that the matching and firing of a production is an all-or-none affair. In the brain, however, there is no evidence for a single, system-wide clock signal capable of sequencing events at the level of simple cognitive operations. For this reason, the architecture we present makes a commitment to distributed timing, in which components at the lowest level time their own operations. Also, since cortical neurons appear to encode information in some form of analog, rather than binary, code, the architecture makes a further commitment to analog representations of the state of a system. These commitments require an approach to duration encoding and sequential processing that is different than the one taken in almost all digital systems.

The architecture we present consists of modular winner-take-all networks composed hierarchically. Each module is layered, with distinct input and output layers, and an intermediate layer that delays propagation from input to output by a controllable amount. In fact, the modules are themselves constructed out of a more basic primitive, a structured column of neural units. Columns are arranged in parallel with lateral inhibitory connections to form modules. Individual units are taken to model the spatiotemporal average firing rate of a population of spiking neurons.

Model neurons

A unit's activation at any moment is represented by a number between 0 and 1. The activation of unit i, $V_i \in [0,1]$, is determined by a standard nonlinear differential equation that is taken to model the firing rate of a population of neurons, possibly averaged over time so that more recent firing contributes more to the average than firing that occurred longer ago:

$$\frac{dV_i}{dt} = -V_i + \frac{1}{1+e^{-\lambda(NetIn_i-\beta_i)}} \qquad (1)$$

where $NetIn_i = \sum_{j=1}^{n} w_{ij}V_j$, and w_{ij} is the synaptic weight on the connection from unit j to unit i (Cohen & Grossberg, 1983). A small random noise term is also often added to V_i. The sigmoid function $f(NetIn_i) = \frac{1}{1+e^{-\lambda(NetIn_i-\beta_i)}}$ forms a curve defining equilibrium values of activation for any given (constant) level of net input. Curves of this type are depicted in Fig. 3.

Self-excitation

Positive feedback within individual units plays an important role in what follows. We define a *self-exciting unit* to be a unit i whose output value V_i is weighted by a nonzero synaptic strength w_{ii} and added to the $NetIn_i$ term of its own activation function. Non-self-exciting units have $w_{ii} = 0$.

We state without proof that the dynamics of self-exciting units receiving constant inputs from other units are completely characterized by the 'self-excitation diagrams' of Fig. 3 (proofs can be found in Simen, (2004)). In plot A, two different activation curves are shown.

Each corresponds to a different level of net excitation received from other units: net excitation results in a leftward shift of the activation function by precisely the amount of the excitation, and inhibition results in a similar rightward shift. The horizontal axis reflects only input to a unit from itself. A stair step trajectory formed between the shifted activation function and the straight line with slope $1/w$ (hereafter referred to as the 'reference line'), where w is the self-excitatory connection strength, determines the equilibrium level of activation ultimately achieved by the unit (assuming no change in external inputs), and the size of the stair steps determines the speed of approach to the equilibrium value (larger steps imply faster approach). As a result, strong and weak self-excitation results in qualitatively different behavior. With weak self-excitation, activation that ramps up or down at a controllable and relatively constant rate is possible. With strong self-excitation, depicted in plot B, approximately all-or-none activation levels and memory in the form of reverberating activation are possible.

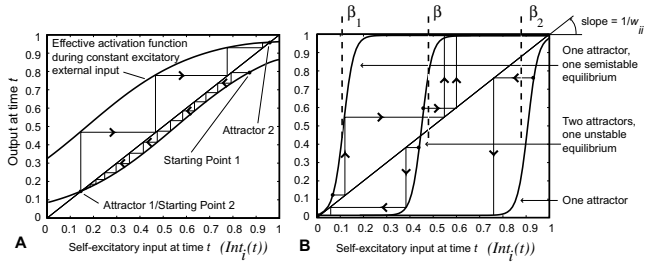

Figure 3: Self-excitation diagrams for units with weak (A) and strong (B) recurrent connections to themselves. The rate at which a unit's activation approaches the nearest attractor is determined by the size of the stair steps depicted.

Decision-making/memory modules

The basic structural primitive of the architecture is a column consisting of a small number of units, but the architecture is easier to approach by beginning with a simple, modular network structure that does not involve columns. This structure, henceforth called simply a 'module', is a fully connected recurrent network in which each unit is connected with a bidirectional connection to every other unit, and in which each unit also excites itself with a strength of 0 or more. The typical mode of operation of a module is as a winner-take-all network, in which one unit becomes highly active and all others become inactive as a result of external input, or 'votes'. Thus modules collect preferences for various decisions and amplify the preference for the most preferred outcome at the expense of those less preferred.

Self-excitation diagrams are useful for module design because they allow a programmer to determine the inter-module connection strengths necessary to create modules that act like AND and OR gates and inverters in digital logic, as well as memory components like latches

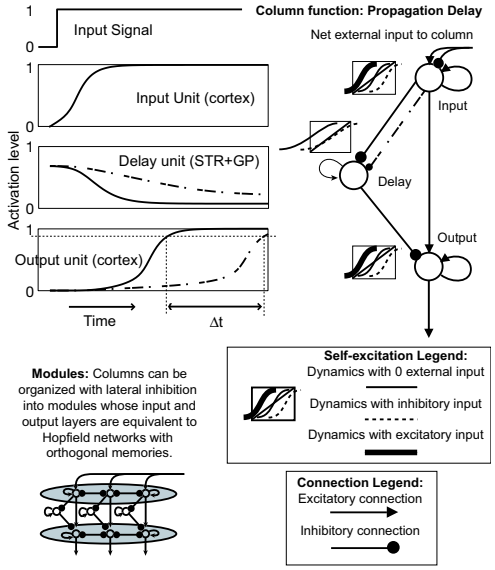

Figure 4: Activation levels in a single column show the effect of weakening the inhibitory strength of the connection from Input to Delay: an increase in propagation delay of size Δt results from slower inactivation of the Delay unit. Modules organized from laterally connected columns are shown at the lower left. Input and Output units model deep and superficial layers of cortex, respectively, and Delay units model the sequence of connections from neurons in the striatum (STR) to the globus pallidus (GP) to the thalamus.

and flip-flops. Most importantly, they allow the setting of connection strengths that cause modules to implement an analogue of production matching and conflict resolution in production systems. Simen (2004) describes this strength-setting recipe in detail.

Model cortical columns

A physical aspect of electronic digital logic components that can have major effects on the performance of a circuit is the internal delay that signals face when propagating through them. In order for high-level descriptions of circuits to work, their components must satisfy a set of timing constraints. The same is true for neural circuits, and a means of adjusting the internal propagation delay within a module to allow timing constraint satisfaction is now presented.

In fact, though, the need for control over propagation is more extensive here than in digital circuit design, because these internal delays will be the sole source of time measurement in the architecture, and timing will be largely asynchronous (without reference to a single, system-wide clock). For this purpose, the architecture elaborates on the module concept with a column structure.

The columnar version of a module is schematically depicted in Fig. 4. Instead of a single, fully connected recurrent network, a module now consists of two identical

copies of such a network. One copy functions as the input interface to the module, and the other functions as the output interface. Each input layer unit, Input, in addition to the lateral inhibition it sends and receives from other input layer units, sends a feedforward excitatory connection to its counterpart Output in the output layer. The Output unit is inhibited by a self-exciting unit, Delay, that also receives inhibitory input from Input. The Delay unit serves to prevent rapid transmission to Output of large jumps in the activation of Input. The rate of transmission from Input to Output is determined by how strongly Input inhibits Delay.

Variable delay characteristics derive from the fact that strong inhibition from Input shifts the Delay unit's effective activation sigmoid to the right. Typically the Delay unit's activation curve sits far enough to the left that it has a single equilibrium value near 1. When shifted to the right, the Delay unit's activation approaches a new equilibrium value near 0. If the rightward shift places the activation curve near to the reference line, this approach will be slow. A sequence of columns imposing a propagation delay on a signal propagating through it can therefore be used as a timer. The effect of weakened Input → Delay inhibition is shown in Fig. 4. This effect is at the heart of the explanation of latency impairments provided here.

Neural model of the Tower of London task

In the model solver, illustrated in Fig. 5, a set of Sensory modules, one for each position of the gameboard, is initialized to patterns encoding the color of a ball at that position, if any, and these representations then persist until reinitialized by changes in the environment. They excite the representations of legal moves in a separate Move module devoted to action representations, and inhibit illegal ones. Attractor dynamics within the Move module results in the selection of a single action for execution, completing the simulation of a simple production of the form: 'if the red ball is in position X, then place it in position Y'.

Goals and subgoals

In the Tower of London solver, one set of winner-take-all modules is dedicated to the representation of externally defined goals and another to internally generated subgoals. Activation in the goal modules biases the competition taking place in the Move module, favoring one column over the others. This biasing is just another form of production, but the if-condition is semantically special: it represents a desired state of the environment. Further, the biasing strength of such a production is insufficient to activate its then-condition without support from some other module, as in the case of the Sensory module just discussed. Technically, this Goal → Move excitation should be considered only a component of a production of the form: 'if Goal is X and Percept is Y, then Do Z'.

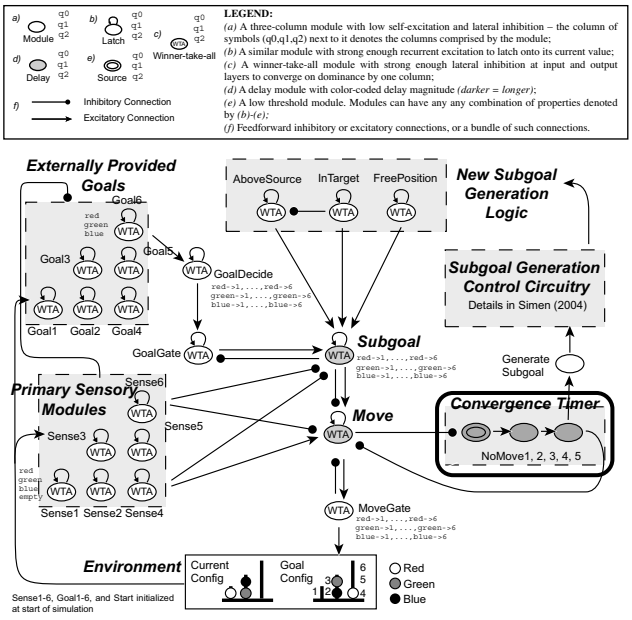

Figure 5: Tower of London model schematic. Each oval is a multi-layer module as depicted in Fig. 4. The highlighted convergence timer component is the only component to make significant use of internal propagation delay. Slowing in this component resulting from weak Input → Delay inhibition is the source of the model's latency impairment.

Convergence timing

The scheme so far described for building problem-solving neural networks addresses critical issues, but it is not obvious that a system of continuous-time winner-take-all modules composed with between-module connections that form closed loops will work properly. The prefrontal/normal model of Polk *et al.* (2002), for example, is actually a hybrid neural/symbolic system. It requires a non-neural component to read the output of a feedforward composition of modules with no cycles. This output emerges in the Move module representing the current action. If a single, clear winner emerges in this network during voting by Sensory and Goal modules, the system takes the prescribed action and then reinitializes several modules in the network to new values. It thereby closes the loop that feeds output information back into the system. However, in the smoothly continuous systems proposed here, such a process cannot be instantaneous: new module values will have a nonzero rise time, and old values will have a nonzero decay time. Thus a timing mechanism is required for ensuring proper 'setup' of inputs for the next cycle of computation.

In many situations, however, an action should never emerge, because no atomic action can achieve the current goal. In such a case, the Move module remains near baseline activation, thereby signaling that a subgoal ought to be generated in order to produce environmental conditions suitable for taking actions to achieve the parent

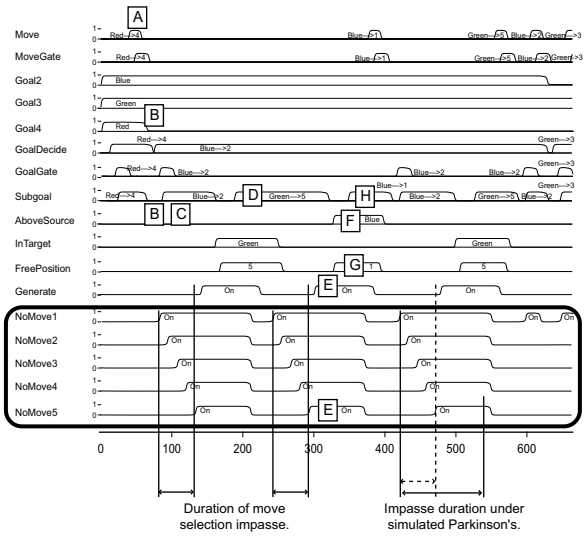

Figure 6: Time course of Output unit activation in most modules of the model during the solution of a five-move-minimum problem, depicted in the Environment panel at the bottom of Fig. 5. Delay between onset of activation in NoMove1 and NoMove5 defines the time window in which a move can be selected before a subgoal is generated. This delay increases as Input → Delay inhibition is weakened, producing the model's latency impairment.

goal. This raises a second question: for how long should an election be allowed to continue? Convergence times within attractor networks (of which winner-take-all networks are a special case) are difficult if not impossible to predict in many cases.

The simple solution presented here is simply to wait for some period of time after one pattern of activation has remained approximately unchanged to declare either a winner or convergence to baseline: the waiting period can be determined by the cost of waiting too long to get a true winner relative to the cost of responding too quickly with a false winner. If we can sample the activation of a module only after a 'safety period' has elapsed, we can reduce the chance of picking a false winner. If we are able to construct a timer which is triggered whenever module activation fails to satisfy the criteria for representing a winner, and which is inactivated by the emergence of a winner, we can produce a convergence-to-baseline detector that is triggered when the timer reaches a threshold duration without being inactivated. This is the function of the Convergence Timer mechanism in Fig. 5.

Model performance

Timecourses of activation in key components of the model are shown in Fig. 6. The problem, shown at the bottom of Fig. 5, requires five moves for solution and therefore requires that some balls be moved to positions other than their final, goal positions. Thus it requires the internal generation of subgoals for efficient solution. The Sense modules, like the Goal modules, are initialized at the beginning of the simulation and excite potentially legal moves. A winner, 'Red to 4' is selected at time point A, and the corresponding unit in MoveGate is caused to rise to threshold, achieving the move and wiping out the move-generating command in Move. At this point, the simulated environment causes an update of the Sense modules, which in turn extinguish any goal or subgoal activation pattern in the Goal system or Subgoal which represent goals to create the current environmental configuration (point B). This allows the next most preferred goal to be retrieved and worked on, as can be seen in Subgoal at point C. At no point is the clock circuit involved.

Now the next goal, 'Blue to 2', which is unachievable, has been selected, and this in turn generates a subgoal to remove an obstacle. Once a subgoal is selected ('Green to 5', since Green is in the target position of the blue ball, at time point D), the first element of the NoMove timer sequence begins to ramp up, and finally maximal activation reaches the last timer in the sequence at time E (this also happens for the previous goal). This activates the Generate module for generating a subgoal. Finally, the subgoal generation logic computes that the ball above the green source ball is blue, at time F, and that the lowest position on a peg which is neither the source nor the target of the goal is position 1 at time G, and Subgoal responds to this voting at time H. The model continues on in this way until eventually solving the problem in 5 moves, as is shown in the sequence of moves selected by the model.

Effect of simulated Parkinson's disease

PD is simulated by reducing the inhibitory effect of Input connections to Delay units within columns. By weakening connections, the propagation delay inherent in any column will be lengthened. For columns in which no propagation delay is required, the effect of this dopamine depletion is expected to be minimal. For columns in which weak Input → Delay connections are needed to model slow propagation delay, the effect is expected to be pronounced. The convergence timer that determines the maximum interval during which computation of a new move can occur is therefore susceptible to simulated dopamine depletion.

In problems which do not require the generation of a subgoal, because all balls can be moved directly to their goal positions, the convergence timer never needs to expire: a move will always emerge well before the time limit. In problems which do require generation of subgoals (some 3 move problems, and by definition, all 4 and 5 move problems), the timer will expire once for each subgoal generated in the simulated dorsolateral PFC.

Fig. 7 shows the interactive effect of simulated dopamine depletion and problem difficulty on the total time to solution of the Tower of London model. Since the model does not store plans, the comparison to the latency data of Owen *et al.*, (1995) is based on the notion that the complete solution of a problem by the model is equivalent to the complete generation of a plan prior to execution by subjects (which is assumed to precede

the first move made by the subject – latency in Owen *et al.* (1995) is therefore the time from presentation of the problem to execution of the first move). The model was run on all possible problems requiring five moves or less, except for problems which would cause difficulty for the simple algorithm implemented by the model. Including modules implementing the heuristics necessary to solve the remaining problems would not be expected to change reaction times significantly, because the modules implementing this logic would not require internal propagation delay (assuming, that is, that these heuristics did not produce a much larger number of moves, but this was not what happened in the hybrid model of Polk *et al.* (2002)).

Simulated dopamine depletion slows the performance of the model in difficult problems. This effect is most pronounced at the five-move-minimum level, and almost nonexistent at the two-move-minimum level. This supralinear increase in latency impairment with increasing problem difficulty is also seen in the performance of PD patients vs. controls. Results therefore support the notion that demands for subgoal representation and maintenance are at the heart of problem solving impairments in prefrontals and PD patients in the Tower of London task.

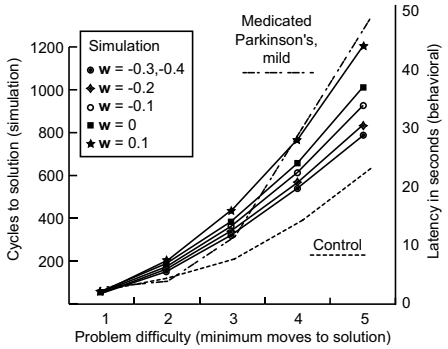

Figure 7: Comparison of average simulation reaction time in solid lines to empirical results in dashed lines (control subjects) and dotted-dashed lines (medicated patients with mild PD). More positive values of **w**, the connection strength from Input to Delay units, corresponds to increasing severity of PD symptoms.

Discussion

The model presented here provides a potential explanation for latency impairments in problem solving by PD patients, in accordance with a previous model that potentially explains prefrontal accuracy impairments (Polk *et al.*, 2002). The structure of the model also hints at the power of a form of neural network modeling that borrows key computational aspects from production systems while adhering to parallel distributed, analog representations of time.

Since dopamine depletion of sufficient severity should also impact the ability of subgoal representations to guide action selection according to this model, accuracy impairments should be expected in severe PD, and this impairment seems supported by the data (Owen *et al.*, 1995). More modeling work is necessary before this expectation can be confidently called a prediction of the model, however. Conversely, since timer circuits are proposed to contain both a cortical component and a component residing in the basal ganglia, the lack of latency impairments in patients with dorsolateral prefrontal lesions remains unexplained. Two possible explanations that do not necessarily conflict with the model are: 1) that timing operations are distributed across a larger area of cortex than just dorslateral PFC, and thus are not disrupted by focal lesions; and 2) since the model only attempts to capture the planning rather than the execution phase of problem solving, it may be that prefrontals show no latency impairment in initiating problem solving despite disrupted internal timing because they impulsively begin execution before a plan is fully developed. Thus they trade off a latency impairment for an accuracy impairment. Future work in which plans are stored, evaluated and finally executed will better draw out the implications of the model presented here.

References

Alexander, G., DeLong, M. & Strick, P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Cohen, M. & Grossberg, S. (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 815–826.

Owen, A., Downes, J., Sahakian, B., Polkey, C., & Robbins, T. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*, 1021–1034.

Owen, A., James, M., Leigh, P., Summers, B., Marsden, C., Quinn, N., Lange, K., & Robbins, T. (1992). Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain*, *115*, 1727–1751.

Owen, A., Sahakian, B., Hodges, J., Summers, B., Polkey, C., & Robbins, T. (1995). Dopamine-dependent frontostriatal planning deficits in early Parkinson's disease. *Nuropsychology*, *9*, 126–140.

Polk, T., Simen, P., Lewis, R. & Freedman, E. (2002). A computational approach to control in complex cognition. *Cognitive Brain Research*, *15*, 71–83.

Robbins, T. & Rogers, R. (2000). Functioning of frontostriatal anatomical 'loops' in mechanisms of cognitive control. In *Control of cognitive processes: Attention and Performance XVIII*. Cambridge: MIT Press.

Shallice, T. (1982). Specific impairments in planning. *Phil. Trans. Royal Soc. London, B*, *298*, 199–209.

Simen, P. (2004). *Neural mechanisms for control in complex cognition.* PhD Thesis, University of Michigan, Ann Arbor.