Striatum and pre-SMA facilitate decision-making under time pressure

Birte U. Forstmann [*], Gilles Dutilh [†], Scott Brown [‡], Jane Neumann [§], D. Yves von Cramon [§], K. Richard Ridderinkhof [*], and Eric-Jan Wagenmakers [†]

[*]University of Amsterdam, Department of Psychology, Amsterdam Center for the Study of Adaptive Control in Brain and Behavior, Amsterdam, The Netherlands,[†]University of Amsterdam, Department of Psychological Methodology, Amsterdam, The Netherlands,[‡]University of Newcastle, School of Psychology, Newcastle, Australia, and [§]Max Planck Institute for Human Cognitive and Brain Sciences, Department of Cognitive Neurology, Leipzig, Germany

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Human decision making almost always takes place under time-pressure. When people are engaged in activities such as shopping, driving, or playing chess, they have to continually balance the demands for fast decisions against the demands for accurate decisions. In the cognitive sciences, this balance is thought to be modulated by a response threshold, the neural substrate of which is currently subject to speculation. In a speeded decision-making experiment, we presented participants with cues that indicated different requirements for response speed. Application of a mathematical model for the behavioral data confirmed that cueing for speed lowered the response threshold. Functional neuroimaging showed that cueing for speed activates the striatum and the pre-SMA, brain structures that are part of a closed-loop motor circuit involved in the preparation of voluntary action plans. Moreover, activation in the striatum is known to release the motor system from global inhibition, thereby facilitating faster but possibly premature actions. Finally, the data show that individual variation in the activation of striatum and pre-SMA is selectively associated with individual variation in the amplitude of the adjustments in the response threshold estimated by the mathematical model. These results demonstrate that when people have to make decisions under time pressure, their striatum and pre-SMA show increased levels of activation.

speed-accuracy tradeoff | basal ganglia | anterior striatum | pre-SMA | Linear Ballistic Accumulator Model

Abbreviations: LBA, linear ballistic accumulator model; fMRI, functional magnetic resonance imaging; pre-SMA, pre-supplementary motor area

Whether buying new shoes, participating in traffic, playing chess, or shooting basketball, one invariably faces the dilemma of when to stop deliberating and make a decision. In many situations, it is maladaptive to ponder over alternative courses of action for a very long time. In basketball, for instance, one has to shoot the ball before a defender can block the shot. On the other hand, decisions taken without sufficient thought may lead to poor results; a shot that is taken too hastily may not go in.

The foregoing example shows that decision-making involves a delicate balance between the competing demands of response speed and choice accuracy, a balance that is usually referred to as the speed-accuracy tradeoff [1]. In the cognitive sciences, this tradeoff is thought to be modulated by a *response threshold* which determines the amount of diagnostic information that is required to make a decision and initiate an action [2, 3]. Because the accumulation of diagnostic information takes time, high response thresholds lead to accurate yet slow decisions, and low response thresholds lead to fast yet error-prone decisions.

The behavioral consequences of the speed–accuracy tradeoff are both profound and predictable, and the tradeoff therefore constitutes one of the most important benchmark findings for formal models of decision-making [4, 5]. In light of its ubiquity and impact, it is surprising that relatively little is known about the neural underpinnings of the speed–accuracy tradeoff (but see [6, 7]). Despite a lack of empirical research, there is a lot of speculation that the basal ganglia may be critical to the speed-accuracy tradeoff [8, 9, 10, 11, 12]. In their default state, the output nuclei of the basal ganglia (i.e., the globus pallidus interna and the substantia nigra pars reticularis) send tonic inhibition to the thalamus, midbrain, and brainstem, preventing the premature execution of any action [13, 14]. When cortical processes start to favor a certain course of action, this leads to activation of input nuclei of the basal ganglia (i.e., the striatum, consisting mainly of putamen and caudate), which, in turn, leads to selective suppression of the output nuclei, releasing the brain from inhibition and allowing the action to be executed [15].

Thus, the basal ganglia are thought to implement a generic action-selection mechanism that releases from inhibition those actions that are desirable, and maintains inhibitory control over all others. The key hypothesis that is shared by recent neuro-computational models of decision-making [8, 9, 10, 11, 12] is that when people have to make decisions under time pressure, the basal ganglia lessen their inhibitory control over the brain in a nonspecific fashion, thereby generally facilitating fast but possibly premature responses.

The goal of this article is to explore the neural correlates of decision-making under time pressure and test the widely held hypothesis that the basal ganglia modulate the speed-accuracy tradeoff. To this end, we experimentally manipulated the speed-accuracy balance in a speeded decision-making task and fitted a mathematical model to the behavioral data. Based on individual differences in the response threshold parameter of this model, functional neuroimaging data revealed that the striatum is involved in the process of setting the response threshold.

Results

In an experiment that consisted of a behavioral session and an fMRI (functional Magnetic Resonance Imaging) session, 19 participants performed a standard "moving dots task" [16]. This task requires a manual response to indicate whether a cloud of moving dots appears to move to the left or to the right. Prior to each stimulus, a pseudo-randomly presented

Reserved for Publication Footnotes

cue indicated the level of speed stress: fast, accurate, or neutral (Figure 1). After each response, participants received feedback consistent with the previously presented cue. In the speed and neutral conditions, participants saw the message "too slow" whenever they exceeded a response time criterion of 450 and 750 msec., respectively. In the neutral and accuracy conditions, participants saw the message "incorrect" whenever they made an incorrect response. This feedback procedure provided additional incentive for participants to adopt different levels of response caution in response to the different cues.

We used the trial-by-trial cuing procedure because this makes it possible to selectively focus on the brain areas that are involved in the setting of the response threshold. A more traditional procedure would cue entire blocks of trials (i.e., an uninterrupted sequence of trials under speed stress, followed by an uninterrupted sequence of trials under accuracy stress), but this would be less appropriate for fMRI, because it would confound the process of setting response thresholds with those involved in processing the stimulus, executing the response, processing feedback, and monitoring performance. For instance, under speed-stress participants make relatively many errors and actively process the stimulus for a relatively short time—blocking the trials would introduce confounds due to differences in processing of error feedback and differences in the extent of active processing of the stimulus.

Behavioral Data. Figure 2A summarizes the behavioral data. We obtained the typical pattern of speed-accuracy results: under instructions to respond more quickly, response time was shorter at the cost of more errors. This effect is most pronounced for the transition to the speed condition, that is, behavioral data from the neutral condition are quite similar to those from the accurate condition, but neither of these conditions are behaviorally similar to the speed condition. Note that accuracy in the speed condition is still well above chance, showing that participants did not resort to fast guessing.

fMRI Data. Figure 3A shows the *conjunction* of two fMRI contrasts—speed vs. accuracy and speed vs. neutral. Whereas in logic a conjunction is defined as an AND between truth statements, in neuroimaging a conjunction refers to activation caused by task X AND by task Y [17]. Thus, our conjunction analysis identifies those brain areas that are active both in the speed vs. accuracy contrast and in the speed vs. neutral contrast. Details of the individual contrasts are provided in Table 1 of the Supporting Information. The conjunction analysis reveals focused, highly reliable ($p < .001$, corrected for multiple comparisons at the cluster-level) activation in the right anterior striatum and the right pre-SMA. Note that the fMRI contrasts are based on the activation elicited by the cue. This means that the focus is entirely on preparatory processes initiated by the presentation of the cue—the later processes initiated by the presentation of the stimulus and the feedback are not the focus of this work and are not reflected in the contrasts. Details of the fMRI procedure and analyses can be found in the Supporting Information.

Thus, our fMRI results suggest that preparation for fast action involves the anterior striatum and the pre-SMA. This conclusion is in line with neuro-anatomical work that shows these brain structures to be part of a closed-loop motor circuit that is involved in preparation and updating of plans for future actions that are under voluntary control [18, 19, 20].

Mathematical Model for Response Speed and Accuracy. These initial conclusions were strengthened when the data were interpreted using a mathematical model for cognitive decision-making, the LBA model [21] (see the method section for additional details). Recall that each time participants were presented with a stimulus, they were required to choose one of two response options. The LBA model represents this choice as a race between two independent accumulators, illustrated in Figure 2B. On each trial, the two accumulators begin with random activation values drawn from independent uniform distributions on $[0, A]$. After the stimulus is presented, activation increases in each accumulator at a rate that depends on the stimulus. For example, activation will generally increase quickly in the accumulator that corresponds to the correct response, but slowly in the accumulator that corresponds to the incorrect response. A response is triggered whenever the first accumulator reaches a fixed response threshold b. Thus, for any decision the observed response time is directly related to the time that the accumulators require to reach the threshold.

For parameter estimation, we used a constrained model in which response threshold b was the only parameter free to vary across cue conditions. We tried many other ways of constraining parameters, but settled on this simple scheme after considering the Bayesian Information Criterion for each design (see Supporting Information for details).

Figure 2C shows that the model fits the data well, using a more complete way of illustrating the data than Figure 2A. Instead of mean response times and error rates, this time we summarize the full response time distributions separately for correct and incorrect responses in each condition. Each distribution is summarized using five quantile estimates (the open circles) which estimate the associated cumulative distribution functions. For each distribution, the slowest (rightmost) symbol represents the 90% quantile—the response time below which 90% of the data fall. The next rightmost symbol represents the 70% quantile, the middle symbol represents the 50% quantile, which is just the median, and the leftmost symbols represent the 30% and 10% quantiles.

Within each panel, the x-axis shows the response time for the quantile estimates, and the y-axis shows the associated proportion of data. This makes the graphs into defective cumulative distribution functions which are commonly used in response time analysis [22]. For example, in the left panel (for data from the accuracy-emphasis condition) the 50% quantile (median) for correct responses was 493 msec. Overall accuracy in that condition was 87%. Half of the *correct* response times fall below the 50% quantile, therefore 43.5% of *all* response times fall below this value, so the data point is plotted at ($x = 493$, $y = .435$).

The quantile estimates support very detailed inspection of the data. For example, in the speed emphasis condition (right panel) the quantile estimates for the incorrect response times were all faster than the corresponding quantile estimates for correct response times. This replicates the usual finding that errors are fast, when speed is stressed [23]. When accuracy was stressed, error response times tended to be about equal to correct response times, but also a little more variable (there was greater separation between the quantile estimates).

Comparison of the data (circles) with model predictions (lines and crosses) in Figure 2C shows that the model fits the data very well. The predicted response probabilities are within 2.1% of the observed values for all conditions, and the predicted response quantiles for correct responses are always within 17 msec of the observed quantiles. Thus, application of the LBA model confirms that the behavioral effects of the experimental manipulation can be entirely accounted for by a change in the response threshold.

Individual Differences. Finally, and crucially, we combined the results from the mathematical modeling and the fMRI measurements [24] and correlated the LBA model parameters derived for each individual with the percent signal change from both the anterior striatum and pre-SMA (see also Supporting Information). Specifically, for each participant and cue condition we calculated the ratio b/A as a measure of response caution. For consistency, we calculated contrasts between parameter estimates in the same way as for the fMRI data. Figure 3B shows a significant negative correlation between the individual changes in the LBA measure for response caution and the individual percent signal change derived from the right anterior striatum. Figure 3C shows the same result for the right pre-SMA. In other words, when put under pressure to respond quickly, some participants adjust their response thresholds more than others. Those participants who have a relatively large decrease in response caution b/A also have a relatively large increase in activation for the right anterior striatum and right pre-SMA. The correlations between response caution b/A and anterior striatum and pre-SMA decrease somewhat when the datum with the highest percent signal change is excluded (i.e, to $r = -.48$ and $r = -.64$, respectively), but the results remain statistically significant (i.e., both p's $< .05$).

We obtained almost identical results when we used a diffusion model [25] to account for the data—specifically, the correlation between EZ boundary separation variable and the hemodynamic response was $r = -.41$ for the anterior striatum, and $r = -.62$ for the pre-SMA (both p's $< .05$). This correspondence shows that our theoretical results generalize across different mathematical models.

The advantages of incorporating in the fMRI analysis the parameters of a mathematical model (instead of more direct summaries of the observed data) are further underscored when one considers the selective nature of the association between individual differences in the hemodynamic response and those in the response caution measure derived from the LBA model. For instance, the association between the hemodynamic response and LBA response caution is much more consistent than the association between the hemodynamic response and mean response time or proportion correct; in fact, for the latter two measures the only significant relation was between proportion correct and activation in the pre-SMA ($r = -.59$, $p = .007$). Also, individual differences in parameters of the LBA model other than response caution (e.g., drift rate) did not correlate with individual differences in the hemodynamic response, further attesting to the selectivity of the association reported in Figures 3B and 3C.

Discussion

Our fMRI analysis showed that when people make decisions under time pressure, this is accompanied by a focused activation in anterior striatum and pre-SMA. The cognitive interpretation of this result was corroborated by fitting the LBA model to the data and by demonstrating an association between the LBA model parameters and individuals' hemodynamic responses. These findings confirm that the striatum is instrumental in adjusting response caution—an assumption from several neuro-computational models of decision-making that has so far evaded experimental scrutiny. Our results are particularly consistent with the model of Lo and Wang [12], who have argued that, in an oculomotor task, time pressure causes an increase in activation in the striatum, which then acts to disinhibit the oculomotor action execution system.

Our results are also partly consistent with recent work by Ivanoff et al. [6] and van Veen et al. [7]. Ivanoff et al. reported that speed emphasis leads to activation in striatum and pre-SMA, as well as in more frontal areas. Van Veen et al. also reported activation of many brain areas, including the striatum, premotor areas of the frontal lobe, and the dorsolateral prefrontal and left parietal cortices. The differences between these results and ours may be due to differences in task, design, computational modeling, and research focus; We tentatively suggest that the relatively broad activation patterns in prefrontal cortex could come about through the use of a suboptimal procedure in which cues precede not just single trials, but entire blocks of trials.

The basal ganglia is a complicated brain structure that is important for reinforcement learning [26], voluntary motor behavior [27], and motor dysfunctions associated with Parkinson's and Huntington's disease [28]. The connections between the pre-SMA and the basal ganglia—more specifically the anterior striatum—render this circuit optimally suitable for modulating action readiness. Accordingly, our work suggests that the striatum, in interaction with the pre-SMA, may also be important in the crucial everyday task of maintaining a balance between fast decisions and accurate decisions.

Materials and Methods

Participants. Twenty healthy volunteers participated for a small monetary reward of 8 euros. All participants signed a consent form prior to the scanning session. All participants had normal or corrected-to-normal vision, and none of them had a history of neurological, major medical, or psychiatric disorders. The data of one participant were excluded from the analysis due to movement. The remaining 19 participants (10 women, M age = 25.5, SD age = 3.08) were all right-handed, as confirmed by the Edinburgh Inventory [29].

Behavioral Task. In the present study we used the moving-dots task, popular in neuroscience and research with primates [16]; for an overview see [30]. Participants were required to decide whether a cloud of dots appeared to move to the left or to the right (Figure1). Out of 120 dots, 60% moved coherently, and 40% moved randomly. Participants indicated their response by pressing one of two spatially compatible buttons with their left or right index finger. A cue ("SN" for schnell, i.e., fast; "NE" for neutral; and "AK" for accurate) instructed participants to adopt different levels of cautiousness on a trial-by-trial basis. The cues were pseudo-randomly intermixed. At the end of each trial, participants received feedback that was dependent on the previously presented cue. In the speed and neutral conditions, participants saw the message "zu langsam" ("too slow") whenever they exceeded a response time criterion of 450 and 750 msec, respectively. In the neutral and accuracy conditions, participants saw the message "falsch" ("incorrect") whenever they made an incorrect response. This feedback procedure provided an additional incentive for participants to adopt different levels of response caution in response to the different cues.

Timing of fMRI Experiment. The timing of the sequence of trials was triggered from the MRI control every 10 seconds. The trials started with a variable oversampling interval of 0, 500, 1000 or 1500 msec to obtain an interpolated temporal resolution of 500 msec. During the variable oversampling interval a fixation cross was presented. Participants were asked to maintain fixation. Then one of the three cues was presented in the middle of the screen for 4800 msec (Figure 1). Cue presentation was followed by a jittered interval between 0 msec to 1500 msec in steps of 500 msec. The imperative stimulus (i.e., the moving dot pattern) was presented for 1500 msec and followed by 350 msec feedback.

The experiment consisted of 240 trials including 24 null events that were pseudo-randomly interspersed. The null events were included to compensate for the overlap of the blood-oxygenation level dependent (BOLD) response between adjacent trials. The experiment lasted about 40 min. Every block started out with two dummy trials which were excluded from further analysis.

Details of the fMRI procedure and analyses can be found in the Supporting Information.

Behavioral Session for the Estimation of Response Threshold. Two days prior to the scanning session, each participant performed the task outside the MRI scanner for about 40 minutes. This yielded sufficient data for the reliable estimation of the response boundary thresholds using the LBA model described below. The trial timing of the task for this behavioral session was modified to maximize the number of observations, i.e., the cue-stimulus interval was set to 500 msec and there was no

variable jitter at the beginning of the trial. Moreover, cue and stimulus were each presented for 1000 msec and there were no null events interspersed. The behavioral session features a total of 840 trials, equally distributed over the three conditions.

Analysis Using a Mathematical Model for Response Speed and Accuracy. We analyzed the data using the LBA model [21].[1] Each time participants were presented with a stimulus, they were required to choose one of two response options – either "left" or "right". The LBA model represents this choice as a race between two independent accumulators, illustrated in Figure 2B. On each trial, the two accumulators begin with random activation values drawn from independent uniform distributions on $[0, A]$. After the stimulus is presented, activation increases in each accumulator at a rate of d units per millisecond, where we call d the "drift rate". The drift rate is a random sample from a normal distribution, with variance s^2 and a mean value that depends on the stimulus (for example, the drift rate for the accumulator that responds "left" will be large when the stimulus strongly suggests that response, and small when the stimulus suggests the other response). A response is triggered whenever the first accumulator reaches a fixed response threshold, b, and the time taken for that response is the time taken to reach the threshold plus a constant offset time t_0.

The predicted response time distributions and associated response probabilities for the LBA can be specified in closed forms [21], and used to calculate likelihood functions when fitting the model to data from the behavioral phase of the experiment.[2] To fit the data, we first removed observations with response times smaller than 250 msec on the grounds that these observations were unlikely to have arisen from the decision process of interest. This trimming resulted in the removal of 105 observations, or only 0.3% of the data. We then estimated response time quantiles corresponding to .1, .3, .5, .7 and .9 cumulative probabilities. The response time quantile corresponding to, say, probability .7 is just that response time below which 70% of the data fall, and these quantiles can be used to succinctly describe response time distributions. We calculated the five quantile estimates separately for each participant in twelve experimental conditions (three response caution conditions crossed with two responses and two stimuli). For fixed model parameters, the probability mass predicted by the LBA model for each inter-quantile bin was computed, and these were combined using the quantile maximum product method [32, 33]. The parameters were then adjusted to maximize the probability product, independently for each participant, using the simplex algorithm [34].

For a decision between two responses, the LBA model appears to have seven parameters: t_0, A, b, as well as means and standard deviations for the drift rate distributions for both left (d_L and s_L) and right (d_R and s_R) responses. For reasons of plausibility and parsimony, we limited these parameters considerably. We constrained the offset time (t_0) and the range of the start point distribution (A) to be fixed across all conditions, and we allowed the response threshold (b) to vary only with response caution. For a given stimulus class (left or right) we used an identical standard deviation for the drift rate distributions for both responses (left and right) and we constrained the means of the drift rate distributions to add to one.[3] This resulted in two free parameters for drift rates – d and s – which we estimated separately for left-moving and right-moving stimuli. These constraints allowed the model to predict the twelve separate distributions of response times using just nine free parameters: t_0, A, b_S, b_A, b_N (for speed, accuracy and neutral conditions), and d_L, s_L, d_R, s_R (for left and right stimuli). We tried many other designs for constraining parameters, such as allowing t_0 to vary with response caution, or fixing the drift rate parameters across stimulus classes. All reasonable designs resulted in similar goodness-of-fit for the model, and in qualitatively consistent results for parameter analyses.

Figure 2C shows that the model fits the data very well. The predicted response probabilities are within 2.1% of the observed values for all conditions. The response quantiles predicted by the model are within 17 msec of the observed quantiles for all distributions associated with correct responses (see also Result Section). The misses are larger for distributions associated with incorrect responses – up to 106 msec – primarily because the data for incorrect responses were fewer and more variable than for correct responses. The model also captures all the important qualitative trends in the data. For example, both correct and incorrect response time distributions are faster (closer to the left) in the speed condition than in the other two response caution conditions. Also, the response accuracy is lower in the speed condition than in the other two conditions (the y-axis probabilities for the two distributions in the speed condition are closer together than in the other conditions).

The LBA model captures response caution through the relative sizes of the response threshold (b) and the upper end of the distribution of starting points (A). When the response threshold is set close to start point distribution (i.e., when $b \approx A$) responses are very fast, but are often incorrect. However, when $b >> A$, the model predicts longer accumulation times and so there is a reduced effect of different starting points for the accumulators. This leads to greater accuracy at the expense of slower responses. As a measure of response caution, we calculated the ratio b/A separately for each participant and separately for the three response caution conditions. The means (and standard errors) for these were 1.66 (0.072), 1.54 (0.069) and 1.17 (0.057) in the accuracy, neutral and speed conditions, respectively. The increase in response caution when moving from speed emphasis to neutral and accuracy emphasis conditions confirms that the experimental condition had the desired effect. The relative similarity of the response caution measures in the neutral and accuracy conditions confirms the conclusion we drew from the descriptive analyses; that participants treated these two conditions similarly. Therefore, and for consistency, when comparing our model analyses with fMRI data, we averaged the estimates from the accuracy and neutral conditions, and made a single contrast—subtracting the response caution measure in the combined conditions from the response caution measure in the speed condition.

ACKNOWLEDGMENTS. We thank Ruud Wetzels for his help in programming the moving dots task, and Mandy Naumann for her assistance with the fMRI measurements. This work was supported by NIH grant R01 MH74457, and by Cognition Pilot, Vidi, and Vici grants from the Netherlands Organization for Scientific Research.

1. **W.A. Wickelgren, Speed-accuracy tradeoff and information processing dynamics, Acta Psychologica, 41 (1977), pp. 67–85.**
2. **R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J.D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks, Psychological Review, 113 (2006), pp. 700–765.**
3. **R. Ratcliff, A theory of memory retrieval, Psychological Review, 85 (1978), pp. 59–108.**
4. **R. Ratcliff and P.L. Smith, A comparison of sequential sampling models for two-choice reaction time, Psychological Review, 111 (2004), pp. 333–367.**
5. **E.J. Wagenmakers, R. Ratcliff, P. Gomez, and G. McKoon, A diffusion model account of criterion shifts in the lexical decision task, Journal of Memory and Language, 58 (2008), pp. 140–159.**
6. **J. Ivanoff, P. Branning, and R. Marois, fMRI evidence for a dual process account of the speed-accuracy tradeoff in decision-making, PLoS One, 3 (2008), e2635.**
7. **V. van Veen, M.K. Krug, and C.S. Carter, The neural and computational basis of controlled speed-accuracy tradeoff during task performance, Journal of Cognitive Neuroscience, in press.**
8. **R. Bogacz and K. Gurney, The basal ganglia and cortex implement optimal decision making between alternative actions, Neural Computation, 19 (2007), pp. 442–477.**
9. **J.W. Brown, D. Bullock, and S. Grossberg, How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades, Neural Networks, 17 (2004), pp. 471–510.**
10. **M.J. Frank, Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making, Neural Networks, 19 (2006), pp. 1120–1136.**
11. **K. Gurney, T.J. Prescott, J.R. Wickens, and P. Redgrave, Computational models of the basal ganglia: From robots to membranes, Trends in Neurosciences, 27 (2004), pp. 453–459.**
12. **C.C. Lo and X.J. Wang, Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks, Nature Neuroscience, 9 (2006), pp. 956–963.**
13. **G. Chevalier, S. Vacher, J.M. Deniau, and M. Desban, Disinhibition as a basic process in the expression of striatal functions. I. The striato-nigral influence on tecto-spinal/tecto-diencephalic neurons, Brain Research, 334 (1985), pp. 215–226.**
14. **J. Deniau and G. Chevalier, Disinhibition as a basic process in the expression of striatal functions. II. The striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus, Brain Research, 334 (1985), pp. 227–233.**
15. **J.W. Mink, The basal ganglia: Focused selection and inhibition of competing motor programs, Progress in Neurobiology, 50 (1996), pp. 381–425.**
16. **K.H. Britten, M.N. Shadlen, W.T. Newsome, and J.A. Movshon, The analysis of visual motion: A comparison of neuronal and psychophysical performance, Journal of Neuroscience, 12 (1992), pp. 4745–4765.**
17. **T. Nichols, M. Brett, J. Andersson, T. Wager, and J.-B. Poline, Valid conjunction inference with the minium statistic, NeuroImage, 25 (2005), pp. 653–660.**
18. **G.E. Alexander and M.D. Crutcher, Functional architecture of basal ganglia circuits: Neural substrates of parallel processing, Trends in Neurosciences, 13 (1990), pp. 266–271.**

[1] The data were also analyzed with the EZ-diffusion model [25], yielding qualitatively identical results.

[2] We performed similar analyses on the entire data set, including data collected in the magnet. The results were similar, but noisier due to increased response variability in the scanner – in particular, there were many anticipatory responses.

[3] The value of 1 was chosen arbitrarily, to enforce a scaling property of the LBA model.

19. M. Inase, H. Tokuno, A. Nambu, T. Akazawa, and M.Takada, Corticostriatal and corticosubthalamic input zones from the presupplementary motor area in the macaque monkey: Comparison with the input zones from the supplementary motor area, Brain Research, 833 (1999), pp. 191–201.
20. K. Shima and J. Tanji, Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements, Journal of Neurophysiology, 80 (1998), pp. 3247–3260.
21. S.D. Brown and A.J. Heathcote, The simplest complete model of choice reaction time: Linear ballistic accumulation, Cognitive Psychology, in press.
22. R. Ratcliff and F. Tuerlinckx, Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability, Psychonomic Bulletin & Review, 9 (2002), pp. 438–481.
23. R. Ratcliff and J.N. Rouder, Modeling response times for two-choice decisions, Psychological Science, 9 (1998), pp. 347–356.
24. B.U. Forstmann, W.P.M. van den Wildenberg, and K.R. Ridderinkhof, Neural mechanisms, temporal dynamics, and individual differences in interference control, Journal of Cognitive Neuroscience, in press.
25. E.J. Wagenmakers, H.J.L. van der Maas, and R.P.P.P. Grasman, An EZ-diffusion model for response time and accuracy, Psychonomic Bulletin & Review, 14 (2007), pp. 3–22.
26. A.M. Graybiel, The basal ganglia: Learning new tricks and loving it, Current Opinion in Neurobiology, 15 (2005), pp. 638–644.
27. J.C. Houk, J.L. Davis, and D.G. Beiser (Eds.), Model of Information Processing in the Basal Ganglia (2nd ed., 1998), Cambridge, MA: MIT Press.
28. T. Wichmann and M.R. DeLong, Functional and pathophysiological models of the basal ganglia. Current Opinion in Neurobiology, 6 (1996), pp. 751–758.
29. R.C. Oldfield, The assessment and analysis of handedness: The Edinburgh Inventory, Neuropsychologia, 9 (1971), pp. 97–113.
30. J.I. Gold and M.N. Shadlen, The neural basis of decision making, Annual Review of Neuroscience, 30 (2007), pp. 535–574.
31. G. Lohmann et al., Lipsia — A new software system for the evaluation of functional magnetic images of the human brain, Computerized Medical Imaging and Graphics, 25 (2001), pp. 449–457.
32. A. Heathcote and S.D. Brown, Reply to Speckman and Rouder: A theoretical basis for QML, Psychonomic Bulletin & Review, 11 (2004), pp. 577–578.
33. A. Heathcote, S.D. Brown, and D.J.K. Mewhort, Quantile maximum likelihood estimation of response time distributions, Psychonomic Bulletin & Review, 9 (2002), pp. 394–401.
34. J.A. Nelder and R. Mead, A simplex algorithm for function minimization, Computer Journal, 7 (1965), pp. 308–313.

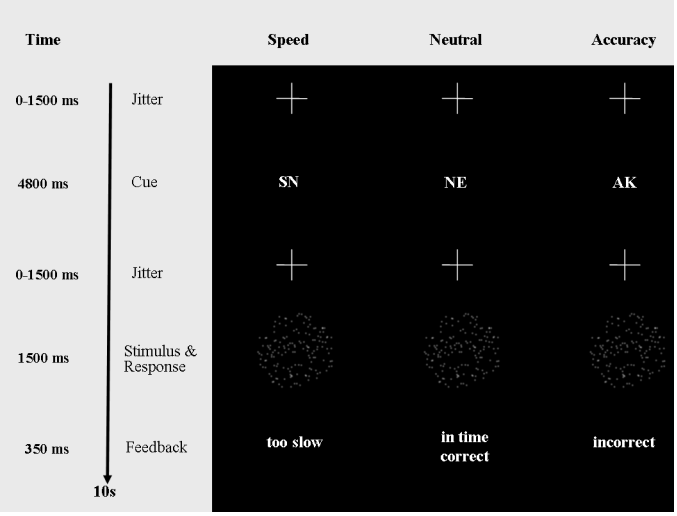

Fig. 1. Paradigm Outline. Moving dots paradigm with cues emphasizing Speed ("SN" for "schnell"), both speed and accuracy, that is, Neutral ("NE"), and Accuracy ("AK" for "akkurat").

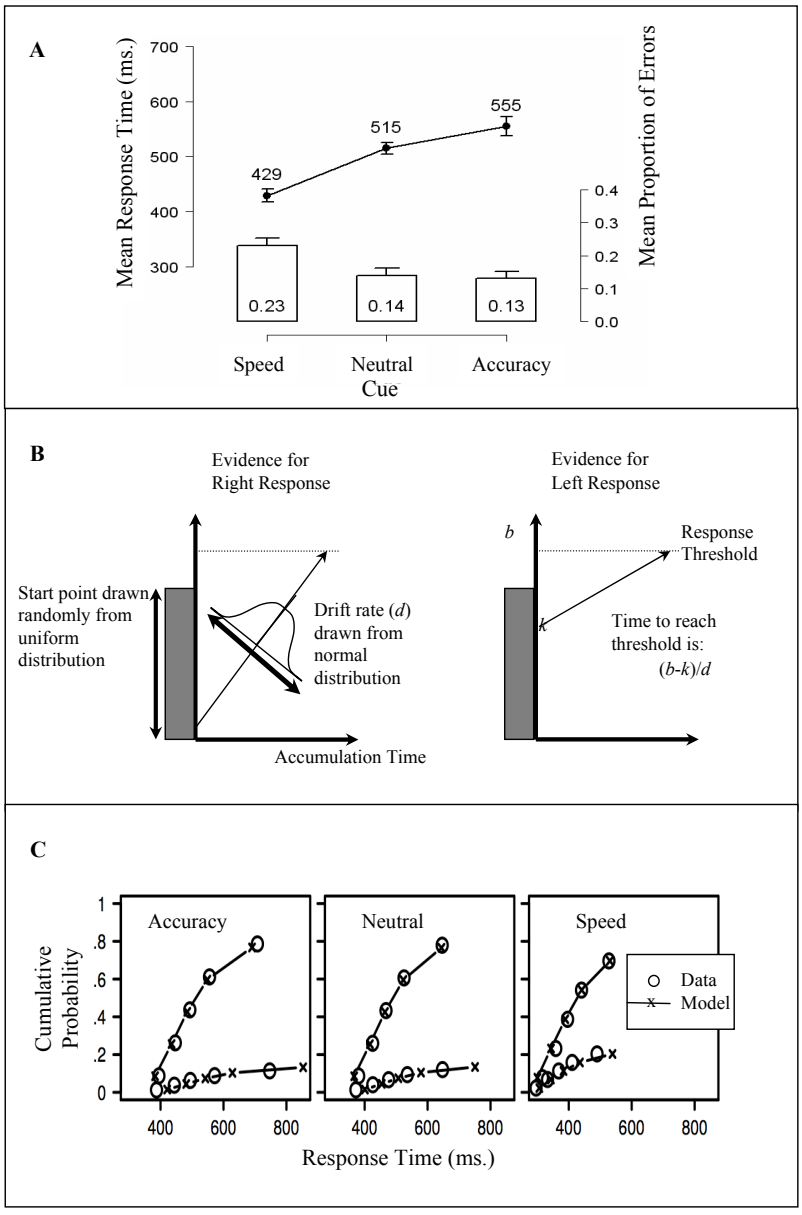

Fig. 2. (A) Behavioral results. Cueing for speed leads to decrease in response time and an increase in errors. **(B) The Linear Ballistic Accumulator model.** In the LBA model, the decision to respond either "left" or "right" is modeled as a race between two accumulators. Activation in each accumulator begins at a random point between zero and A and increases with time. The rate of increase is random from trial to trial, but is (on average) faster for the accumulator whose associated response matches the stimulus. A response is given by whichever accumulator first reaches the threshold b, and the predicted response time depends on the time taken to reach that threshold. **(C) Model fit.** Quantiles estimated from data (circles) and predicted by the LBA model (crosses with lines). The three panels show data from three different response caution conditions. Within each panel, the upper lines and symbols show quantile estimates for correct responses, and the lower set for incorrect responses. The data and model predictions were averaged across participants and across left vs. right stimuli.

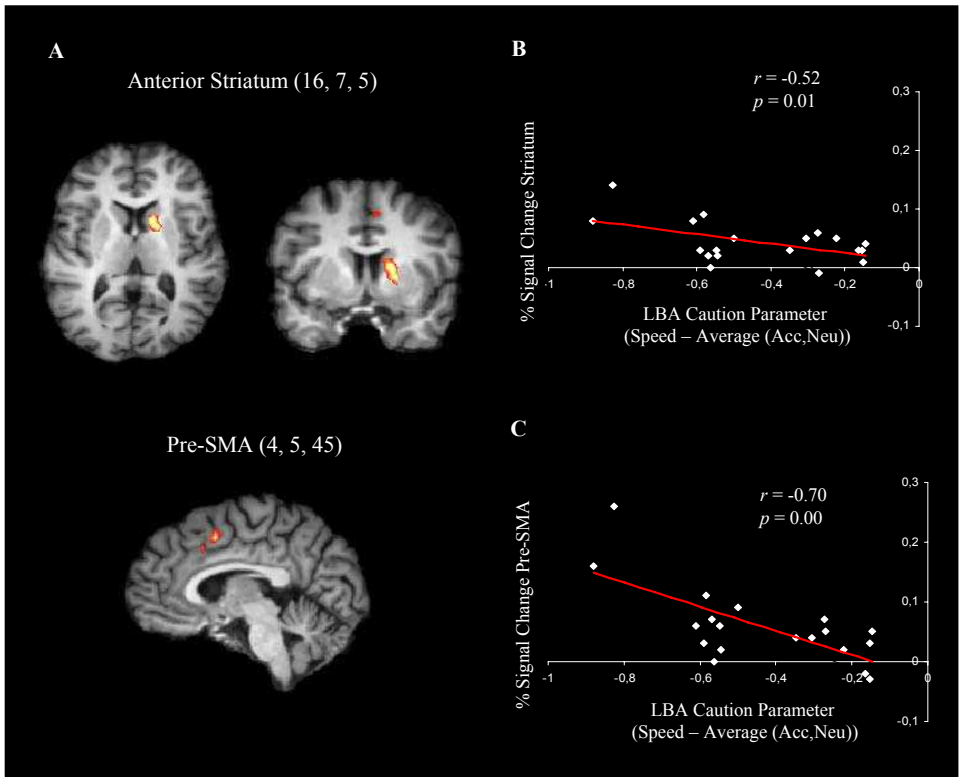

Fig. 3. Conjunction and correlation analyses. Activation maps averaged over 19 participants mapped onto an individual brain. Red labels indicate positive z values. Coordinates are given in Talairach space. (A) Activation elicited in the conjunction analysis of both Speed vs. Accuracy and Speed vs. Neutral. (B) Association between the individual percent signal changes derived from the right anterior striatum and the individual changes in the LBA measure for response caution. (C) Association between the individual percent signal changes derived from the right pre-SMA and the individual changes in the LBA measure for response caution.