Capacity sharing between concurrent movement plans in the frontal eye field during the planning of sequential saccades

Debaleena Basu[1*], Naveen Sendhilnathan[2], Aditya Murthy[1]

[1]Centre for Neuroscience, Indian Institute of Science, Bangalore, Karnataka 560012, India

[2]Department of Neuroscience, Columbia University in the City of New York, NY 10027, USA

*Corresponding author: Dr. Debaleena Basu; e-mail: basu.debaleena@gmail.com

Number of pages: 35

Figures: 7

Supplementary figures: 5

Summary:

A ramp to threshold activity of frontal Eye Field (FEF) movement-related neurons best explains the reaction time of single saccades. How such activity is modulated by a concurrent saccade plan is not known. Borrowing from psychological theories of capacity sharing that are designed to explain the concurrent planning of two decisions, we show that processing bottlenecks are brought about by decreasing the growth rate and increasing the threshold of saccade-related activity. Further, rate perturbation affects both saccade plans, indicating a capacity-sharing mechanism of processing bottlenecks, where both the saccade plans compete for processing capacity. To understand how capacity sharing was neurally instantiated, we designed a model in which movement fields that instantiate two saccade plans, mutually inhibit each other. In addition to predicting the greater reaction times for both saccades and changes in growth rate and threshold activity observed experimentally, we observed a greater separation of the two neural trajectories in neural state space, which was verified experimentally. Finally, we show that in contrast to movement related neurons, visual activity in FEF neurons are not affected by the presence of multiple saccade targets, indicating that inhibition amongst movement-related neurons instantiates capacity sharing. Taken together, we show how psychology-inspired models of capacity sharing can be mapped onto neural responses to understand the control of rapid saccade sequences.

Introduction:

Processing bottlenecks have been investigated widely using dual-task paradigms (Pashler, 1994; Ruthruff, Pashler, & Klaassen, 2001), which involve making two independent responses (for example, an eye and a hand movement) to two stimuli (i.e., eye target and hand target) appearing in close succession. The results universally show that as the temporal gap between the targets (TSD; target step delay) decreases, the latency of the response to the second stimulus increases markedly, as if the brain inherently cannot process two simple decisions at the same time. The delaying of the second response is referred to as the dual-task interference, and is thought to reflect bottlenecks in information-processing stages that come into play when two motor plans are proceeding to completion simultaneously (Marois & Ivanoff, 2005; Pashler, 1994). While dual-tasks typically constitute responses with two separate effectors, a sequence of two saccades may be considered as a single-effector, two-response system and can be studied using the dual-task framework. The typical pattern of increase in the second response latency with decrease in TSD, characterizing processing bottlenecks, has also been observed in sequential saccades (Ray, Bhutani, & Murthy, 2012; Ray, Schall, & Murthy, 2004).

Various theoretical frameworks have been proposed to explain how simultaneous or closely spaced action plans interfere with each other. Single-channel bottleneck models propose that a central, decision-making stage is critical for each action plan, and that the central stages of multiple plans can only proceed serially, and cannot be 'co-active' (Pashler, 1994; Ruthruff et al., 2001; A. Welford, 1967; A. T. Welford, 1952). Thus, the central stage imposes a bottleneck, through which only one plan can pass at a time. In the case of a sequence of two saccades, the first plan is likely to reach the central stage first, and thus the S2 (saccade 2) plan has to 'wait' till central processing of the first is over (**Fig. 1A**). In the scheme of single-channel bottleneck models, the second plan shows the hallmark of processing bottlenecks: increase in latencies with decrease in TSD, whilst the S1 (saccade 1) latencies stay unaffected. In contrast, capacity-sharing models argue that the decision-making stages of both plans can proceed in parallel, albeit the processing capacity is shared differently. The concept of the brain's 'capacity' corresponds to the brain's general information processing capabilities (Broadbent, 1971; Gopher & Navon, 1980; Kahneman, 1973; McLeod, 1977), independent of task type. This limited capacity can be shared across tasks and the capacity-sharing theory of dual-task interference proposes that smaller the proportion of capacity allocated to a task, the longer the task plan takes to reach completion (Navon & Miller, 2002; Tombu & Jolicœur, 2003). The capacity-sharing models predict that because of its temporal precedence, the S1 plan will get

the major share of the capacity and the S2 plan will get a smaller fraction, thus delaying the onset of the second response (**Fig. 1B**). However, unlike single-channel bottleneck models, where plan 1 may be assumed to get 100% of the capacity, capacity-sharing models predict that the latencies of the S1 will also increase as it only gets a part of the full available capacity. Another difference is that the capacity-sharing models propose that the central stages of the 1st and 2nd tasks can proceed in parallel, while single-channel bottleneck models postulate that they are processed serially, one after the other. While a plethora of behavioral studies have investigated processing bottlenecks in dual-task studies, similar work in the domain of sequential saccades has been lacking. Further, the neural mechanisms at the level of single unit activity underlying processing bottlenecks in both dual tasks and sequential saccades remain unknown.

We recorded from frontal eye field (FEF) neurons to search for signatures of processing bottlenecks as macaque monkeys performed a modified version of the double-step saccade task (Westheimer, 1954), the FOLLOW task (**Fig. 1C**). FEF neurons are functionally divided into two main categories based on whether the firing activity is evoked in response to stimulus onset (visual neurons) and/or whether the activity leads to saccade initiation (visuomovement neurons/movement neurons (Bruce & Goldberg, 1985; Hanes & Schall, 1996). The activity of FEF movement neurons closely follows accumulator dynamics and represents a decision-making stage, which has been found to be capacity-limited in computational models (Hanes & Schall, 1996; Ray et al., 2012; Sigman & Dehaene, 2005). Further, FEF is a higher-order control center for goal-directed saccadic planning(Sendhilnathan, Basu, Goldberg, Schall, & Murthy, 2019; Sendhilnathan, Basu, & Murthy, 2017, 2020). Finally, FEF movement neurons encode two saccade plans in parallel (Basu and Murthy, 2019). Thus, FEF is the candidate region to study the neural imprints of processing bottlenecks, or the limits to parallel processing during the making of rapid sequential saccades.

Results:

Two monkeys, a *Macaca radiata* (J) and a *Macaca mulatta* (G) performed a sequential saccade 'FOLLOW' task (**Fig. 1C**; see methods), where the majority (70%) of the trials were 'step trials' in which they had to perform a rapid sequence of saccades to two targets (T1 and T2) in the order of their presentation. The remaining 30% of the trials were 'no-step' trials, wherein a single visual target was presented and the monkeys had to make a single saccade to

it. The two types of trials were randomly interleaved. The temporal gap (target step delay or TSD) between the T1 and T2 in step trials was randomly chosen among 17 ms, 83 ms, and 150 ms (Basu & Murthy, 2019).

Behavioral evidence of processing bottlenecks during sequential saccades

Behavioral data from a majority of the sessions (n = 44) exhibited the main feature of processing bottlenecks: an increase in the mean S2 RT with a decrease in TSD (**Fig. 1D**; Kruskal-Wallis, χ^2 (2, 165) = 145.27, $p < .001$, $\eta^2 = 0.87$). Post-hoc comparisons using Tukey's HSD criterion showed that the mean RTs at all the TSDs were significantly different from each other. The first saccade latency significantly slowed down across TSD (although the effect was much smaller than RT2) indicating that the S1 plan was not inherently free from processing bottlenecks, suggesting that a capacity sharing mechanism is employed (**Fig. 1D**; Kruskal-Wallis, χ^2 (2, 165) = 39.34, $p < .001$, $\eta^2 = 0.2356$). Post-hoc comparisons using Tukey's HSD criterion showed that the mean RT1 at the shortest TSD of 16 ms was significantly more than the longest TSD. The increase in saccade latencies with decreasing TSD corroborated with previously well-established evidence of processing bottlenecks in concurrent action planning. The remaining 25 sessions also showed that the typical elongation of the second saccade RT with decrease in TSD (**Fig. 1D**; Kruskal-Wallis, χ^2 (2, 81) = 50.04, $p < .001$, $\eta^2 = 0.6029$) indicating processing bottlenecks. Interestingly, the first saccade RT for these sessions was significantly increased for the longest TSD (150 ms) compared to the short and medium TSD (**Fig. 1D**; Kruskal-Wallis, χ^2 (2, 81) = 39.49, $p < .001$, $\eta^2 = 0.4035$).

Movement-related activity during single saccades

Previous work has shown that the pattern of activity of FEF movement neurons are correlated with the stochastic accumulation processes widely used in computational models of saccadic RTs (Boucher, Palmeri, Logan, & Schall, 2007; Hanes & Schall, 1996; Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007; Woodman, Kang, Thompson, & Schall, 2008) and are directly linked to saccade initiation times (Huerta, Krubitzer, & Kaas, 1986; Langer & Kaneko, 1990; Segraves, 1992). Since reaction time lengthening is the main behavioral evidence of processing bottlenecks, movement neurons are well projected to carry neural correlates of the same. To confirm whether movement-related activity in FEF adheres to an accumulation-to-threshold model of reaction time, no-step single-saccade trials were divided into fast, medium, and slow RT groups and parameters of accumulator models were measured from movement activity (**Fig. 2A**). RT grouping was obtained by partitioning RTs in each

135 session using the mean RT of that session (see methods). The main parameters of accumulator
136 models, i.e., baseline, onset, growth rate, and threshold activity were measured for the three
137 RT conditions (fast, medium, slow), for each neuron (**Fig. 2C-F;** see methods). Consistent with
138 the earlier studies(Hanes & Schall, 1996), adjustments in the rate of growth of activity of the
139 movement neuron population predicted RTs in the no-step trials: across the movement neuron
140 population, the slope of the growth rate was significantly different from zero (Z_{rate} = -4.27, p
141 < .001; **Fig. 2E**). Further, the slopes for the rate of activity growth were negative, indicating
142 that fast RTs were preceded by a steeper rate of growth of movement activity and vice versa.
143 While the growth rate varied with reaction time, the threshold did not (Z_{rate} = -0.98, p = .323;
144 **Fig. 2F**), corroborating with the accumulation-threshold model of reaction time. The slope
145 distributions of other accumulator measures like baseline, and onset, were not statistically
146 significant from zero ($Z_{baseline}$ = -2.04, p = 0.05; Z_{onset} = 1.92, p = 0.054).
147
148 **Processing bottlenecks underlie the representation of the S2 plan in FEF movement**
149 **neurons**
150 Movement-related neurons are an excellent candidate to encode the neural correlates of
151 RT lengthening, which is the main behavioral evidence of processing bottlenecks. Based on
152 the mapping between accumulator models and movement neuron dynamics, four possible
153 hypotheses (**Fig. 2A**) can explain how the activity of FEF movement neurons coding for S2
154 might bring about the systematic increase of the S2 RT with decrease in TSD that characterizes
155 processing bottlenecks. The lengthening of RT may be due to (1) lowering of the baseline firing
156 rate with shorter TSDs (2) delaying of the onset of the activity related to the S2 with shorter
157 TSDs (3) reduced growth rate of the activity with shorter TSDs (4) and an increase of the
158 saccade threshold firing rate with larger TSDs. To assess which of the possibilities explain the
159 increase in S2 RT, we analyzed the neural activity in trials where the S2 was made into the RF
160 (RFin trials; see methods) for all three TSDs (**Fig. 3A**). Across the population, the rate of
161 evidence accumulation slowed down from long to short TSD, and the activity ramped up to a
162 higher firing rate threshold.
163 We measured each of the four accumulator parameters: baseline, onset, rate, and
164 threshold (average across trials of the same TSD) for the three TSD conditions, for each neuron
165 (**Fig. 3B;** see methods; see **Fig. S1** for example movement neuron) using linear regression. The
166 slopes from all the movement neurons were compared using a Wilcoxon signed-rank test.
167 Across the movement neuron population, the slope of the rate, and the threshold, as a function
168 of TSD was significantly different from zero (Z_{rate} = 4.27, p < .001; $Z_{threshold}$ = -2.67, p < 0.01;

Fig. 3B). Further, the slopes for the rate of activity growth were positive, indicating that the rate of activity grew faster at longer TSDs, where presumably the effect of processing bottlenecks was the least among the three TSD conditions. Threshold slopes were significantly negative, indicating that as the TSD increased, the threshold required for initiation of the S2 was reduced at the population level. However, the slope distributions of other accumulator measures like baseline, and onset, were not statistically significant from zero ($Z_{baseline} = -0.62$, $p = 0.53$; $Z_{onset} = 0.86$, $p = 0.17$). Thus, processing bottlenecks at the level of FEF movement neurons were characterized by multifaceted adjustments in the rate and threshold of the activity related to S2.

FEF movement neurons display capacity sharing during encoding of two overlapping saccade plans

While the classical evidence of processing bottlenecks is indexed by the increase in the latency of the S2, the S1 latency may also be affected according to the capacity-sharing scheme of processing bottlenecks (**Fig. 1B**). Since the single-channel bottleneck model and the capacity-sharing model differ in their predictions for the S1 plan, we tested whether movement-related activity encoding S1 remained unchanged, or changed systematically, across TSDs, as would be expected in the single-channel bottleneck scheme, or the capacity-sharing hypothesis, respectively. To address this issue, we performed the same analyses as before but for the condition in which S1 was made towards the RF (RFout trials; **Fig 4A**). At the population level, rate perturbation occurred with decrease in TSD in the first plan, mirroring the modulation observed for the second plan (Wilcoxon signed-rank test for slopes of rates, $Z_{rate} = 3.62$, $p < .001$; **Fig 4A; Fig S1**). However, unlike the second plan, threshold activity did not show a significantly decreasing relation with TSD ($Z_{threshold} = 1.16$, $p = 0.25$;). Slope distributions of other accumulator measures like baseline, and onset, were not statistically significant from zero ($Z_{baseline} = 0.85$, $p = 0.39$; $Z_{onset} = 1.04$, $p = 0.29$; **Fig 4B**). Thus, rate perturbation constituted a major mechanism through which the ramp-up of activity of FEF movement neurons was controlled during parallel planning of sequential saccades. However, in the short TSD condition (17 ms), the population activity did not rise linearly with a reduced rate, but rather showed a temporary slowing down before again resuming the ramp-up.

200 **State space dynamics and inhibitory control may enable capacity sharing during**
201 **sequential saccade planning**

202 To gain deeper insights into neural mechanisms underlying capacity sharing, we studied
203 the dynamics underlying the trajectory of neural activity in FEF. More specifically, we
204 hypothesized that at lower TSDs, the mutual inhibition between the two competing plans forces
205 the neural trajectories to occupy different subspaces so that the plans could evolve in parallel.
206 To test this, we performed a principal component analysis (PCA) separately for the population
207 neural activity (for saccades into RF) aligned to T1 and T2 for each of the three TSDs (**Fig**
208 **5A**). PCA is a commonly used unsupervised learning algorithm to extract the latent information
209 from the data (see methods). This method allows us to look at the high dimensional FEF
210 population neural activity in a much lower dimension that captures the maximum variance of
211 the population.

212 At least ~ 7-8 PCs were required to explain >99% of the variance for any of the six
213 conditions (three TSDs, two plans; although there was a trend of fewer PCs explaining more
214 variance as TSD increased). However, the top three PCs explain >90% variance (**Fig 5B**).
215 Therefore, we visualized a 'state-space trajectory' by plotting the top three PCs versus one
216 another (**Fig 5C**). Each point on the trajectory indicates the neural state at each time point. The
217 neural trajectories significantly differed between the planning of S1 and S2 for the shortest
218 TSD but became more similar as the TSD increased (**Fig 5D)**. We also confirmed that these
219 differences were related to the TSDs and not to differences in saccade kinematics, which were
220 similar across TSDs for S1 and S2 (**Fig. S2**).

221 If planning for S1 and S2 compete for the same shared space due to limited capacity,
222 we should expect the neural trajectories to span different subspaces at shorter TSDs and span
223 the same subspace at higher TSDs. In contrast, the bottleneck hypothesis, predicts that the
224 subspaces corresponding to the planning of S1 and S2 to be completely orthogonal, implying
225 that the S2 plan to be completely dormant until plan 1 is completed. To test between these
226 alternatives, we quantified the degree of overlap between the subspaces spanned by these neural
227 trajectories (see methods). At the shortest TSD, the magnitude of overlap between the signals
228 for planning of S1 and S2 was 47% and this increased as the TSD increased from medium
229 (84%) to long TSDs (92%; **Fig 4D**). This result also held true for all saccade directions (**Fig**
230 **S3**).

231 Next, we wanted to investigate if the difference in neural subspace overlap among
232 different TSDs was indeed due to differences in the mutual inhibition of plans co-evolving in

233 parallel. To address this issue, we performed two sets of simulations (see methods; **Fig S4A-**
234 **F**) using accumulator framework (**Fig S4H**). For each of the two sets, we simulated 40 neurons
235 with 900 trials per neuron (with three types of TSD trials) using a firing rate model to
236 approximately match the statistical power of our experimental dataset (see methods). We
237 constructed an inhibition function such that the magnitude of mutual inhibition inversely varied
238 with TSD (see methods; **Fig S4G**). The resulting simulated neural activities (with mutual
239 inhibition) resembled the neural data (**Fig 6A-B; Fig S4K-L**). Here, ~ 7-8 PCs were required
240 to explain >99% of the variance for the shortest TSD and fewer (~5-6) PCs were required to
241 explain >99% of the variance for the longest TSD (**Fig 6C**). The neural trajectories significantly
242 differed for short TSD but were similar for longer TSDs (**Fig 6D**) and the degree of subspace
243 overlap between the two plans increased with TSD, consistent with the structure present in the
244 neural data (**Fig 6E**). In the next set of simulations, we did not introduce any mutual inhibition
245 factor into the simulations (**Fig 6F**). Here, for each TSD, the activities for both plans evolved
246 in parallel and were linearly summated over time. The resulting simulated neural activities were
247 similar for both plans and for all three TSDs (**Fig 6G; Fig S4M**). Very few (~3) PCs explained
248 >99% of the variance (**Fig 6H**). The neural trajectories were not significantly different between
249 planning of S1 and S2 for any of the TSDs (**Fig 6I)** as the subspace overlap was 98% between
250 any pair of plans (**Fig 6J**). Furthermore, the simulations with mutual inhibition reproduced the
251 empirical neural findings during plan 1 and plan 2 but the simulations without mutual inhibition
252 did not (**Fig S5**).

253 **Processing bottlenecks in FEF visually related neurons**

254 Previous studies have reported a separation between the visual and motor processing of
255 FEF neurons with only motor processing affecting RT in perceptually simple tasks (Sato,
256 Murthy, Thompson, & Schall, 2001; Thompson, Bichot, & Schall, 1997; Woodman et al.,
257 2008). Thus, it is plausible that the responses of visual neurons are not gated by inhibitory
258 bottlenecks. Here, we tested this notion using visual target-related activity during the visual
259 epoch for purely visual (**Fig 7A**) and visuomovement neurons (**Fig 7B**).
260 We analyzed the average target-related response in the 200 ms window for each neuron
261 to identify signatures of processing bottlenecks (taking a longer epoch risked incorporating
262 movement-related activity from visuomovement neurons). If target selection is capacity-
263 limited, then presumably, neural responses encoding saccade targets appearing in close
264 succession will be inhibited, either due to single channel bottleneck (only second target

response gets affected) or due to capacity sharing (both first and second target responses get affected). In contrast to movement-related activity, the average activity in the target-related period did not vary with TSD (Kruskalwallis: χ^2 (2, 129) = 0.47, p = .79 (S1); χ^2 (2, 124) = 0.06, p = .97 (S2)) for both saccade plans, suggesting that the visual processing stage is pre-bottleneck, at least of a perceptually simple task like the FOLLOW task.

Discussion

In this study, we explored the limits of parallel processing engaged in saccade sequence planning. Processing bottlenecks were found within FEF, the mechanisms being rate perturbation and threshold modulation in the movement neuron population. In contrast, visual activity did not reflect any major modulation that could be considered a significant bottleneck. Additionally, we found evidence of processing bottlenecks for both S1 and S2 motor plans, suggesting that limited processing was a consequence of capacity sharing between co-activated movement plans. The notion of such shared and limited processing was also revealed in the state space dynamics of FEF movement activity, which showed a potential role for inhibitory control that gated access of concurrent motor plans to a planning subspace. The major results are discussed and interpreted in the following sections.

Processing bottlenecks in sequential saccade planning

Processing bottlenecks and parallel programming represent functionally antithetical processes, and yet both are essential for optimal saccadic behavior. While parallel programming allows for rapid execution of a saccade sequence, processing bottlenecks are likely to arise to check unbridled parallel programming of motor plans as failure to control it might lead to errors like averaged saccades or incorrect order of execution of a saccade sequence (Bhutani, Ray, & Murthy, 2012; Coëffé & O'regan, 1987; Findlay, 1982; Ray et al., 2012; Viviani & Swensson, 1982; Zambarbieri, Schmid, & Ventre, 1987). In the context of the current study, we tested whether a single-channel bottleneck (Pashler, 1994) or a capacity-sharing bottleneck (Kahneman, 1973) best explained our reaction time data since behavioral evidence of both the models have been found in dual-task paradigms (Arnell & Duncan, 2002; Navon & Miller, 2002; Pashler, 1994). In contrast to a strict dichotomy, we found that even though a majority of behavioral sessions reflected capacity sharing in our study (i.e. RT1 increased from high to low TSD), a significant fraction of sessions (25/69), did show evidence of bottleneck processing. Be that as it may, the capacity-sharing model still suffices to explain the pattern of RTs in those sessions, if the first plan capacity is kept at 100%; i.e., it is not

shared. Thus, whereas the single-channel bottleneck model can be derived from the capacity sharing model, the reverse isn't true. As to why capacity sharing should change between sessions, the varying latency measures with TSD may reflect the functioning of executive control that varies with performance of the monkeys from session to session. The flexibility of capacity sharing also extends to cognitive control: if greater priority or 'processing capacity' is given to the second saccade plan over the first, the order of saccades may be reversed or redirect behavior can be elicited (Bhutani et al., 2012; Ray et al., 2012). Nevertheless, even though behavioral evidence of capacity sharing, was at times ambiguous, the neural correlates of capacity sharing were more prominently observed. Such a dissociation between behavior and neural activity is interesting and suggests a form of compensatory redundancy in the activity of FEF movements neurons.

Inhibitory control underlying processing bottlenecks

We tested whether such mutually inhibitory accumulators encoding distinct saccade plans can also instantiate a more nuanced form of inhibitory control wherein both the saccadic eye movements are executed, albeit with greater delays, especially for the second saccade. Modelling such a response required two important additional conditions: the first condition, required that the inhibition be asymmetric, being greater for the first saccade plan than the second saccade plan, which manifest as greater capacity and faster information processing for the former compared to the latter. Such an asymmetry is a natural consequence of the temporal delay allowing for greater activity in the first saccade to inhibit the second saccade; the second condition required an inhibitory kernel that decreased with target step delay, such that inhibition from the first accumulator to be greater at shorter delays despite being the level of activity in the accumulator being lesser compared to what it would be at larger target step delays. Such an inhibitory kernel is necessary to match the observed behavioral data of greater second saccade RTs (**Fig S4**) as well as the neural data which showed greater interference for the second saccade motor plans at the shorter TSDs (**Fig S5**). Interestingly, using a dynamical systems approach under the assumption of stationarity of noise across trials (Elsayed, Lara, Kaufman, Churchland, & Cunningham, 2016), this model of inhibitory control could be also shown to act as a "queue", in which non-orthogonal neural spaces can allow simultaneously allow parallel processing but yet temporarily slow the processing of the second saccade. We believe that the ability of such inhibition to reconfigure the neural space may reflect the nonlinear effects of inhibition on the pattern of activity representing accumulator activity that plan the saccades.

The simplest and most parsimonious explanation for the location of such a bottleneck is at the level of FEF via mutual inhibition (Ray, Pouget, & Schall, 2009) of competing motor plans developing in the FEF. This type of inhibitory gating can be brought about by inhibitory interneurons within the FEF (Markram et al., 2004; Somogyi, 1977). Although such a form of inhibition is intuitive and can be readily implemented within the proposed frameworks described for decision-making circuits (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff & Smith, 2004), implementing an inhibitory kernel that decreases with increasing TSD cannot be easily implemented in a straightforward manner by mutually inhibitory accumulators. Furthermore, using an identical task, our previous work has shown that the BG is causally involved in the conversion of parallel movement plans into sequential behavior (Bhutani et al., 2013). Inactivation of the BG in monkeys with muscimol or impairment of the BG in Parkinson's disease patients resulted in a significantly greater extent of saccadic errors that develop due to unchecked parallel programming leading to a 'collision' of saccade plans. The results of both these studies can be reconciled by the fact that FEF and BG share a closed connection through the cortico-BG-thalamo-cortical loop wherein the thalamus, a major relay center, receives projections from BG output nuclei, and in turn projects to multiple cortical regions, including the FEF, which are again routed to the input nuclei of BG (Alexander, DeLong, & Strick, 1986; Middleton & Strick, 2000; Parent & Hazrati, 1995a, 1995b). Thus , the origin of the bottleneck could also be in the well-established inhibitory control circuitry of the basal ganglia (Hikosaka, Takikawa, & Kawagoe, 2000) and then re-routed to the FEF through the basal ganglia-thalamo-cortical loop (Goldman-Rakic & Porrino, 1985), which then manifests it various adjustments of movement-related neuronal activity. Thus, in addition to local inhibitory interactions within the FEF, inhibitory interactions within the BG are postulated to contribute to processing bottlenecks that we observed. More specifically, we speculate a role of the indirect pathway that has two consecutive inhibitory circuits from the striatum to the GPe, and from the GPe to the GPi, in instantiating the inhibitory kernel that we modelled to explain the data.

Neural representations of processing bottlenecks within FEF

Our results are consistent within a framework in which decision-making and motor planning are instantiated as stochastic accumulators integrating noisy sensory evidence from a baseline level over time; a response is produced when the accumulated evidence reaches a 'decision' threshold (Gold & Shadlen, 2007; Nosofsky & Palmeri, 1997; Ramakrishnan &

366 Murthy, 2013; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Smith & Van Zandt, 2000;
367 Usher & McClelland, 2001). While the accumulator framework has been used to understand
368 dual-task interference as well (Sigman & Dehaene, 2005; Zylberberg, Ouellette, Sigman, &
369 Roelfsema, 2012), this is the first study to reveal the neurophysiological correlates of such
370 processing bottlenecks in the context of sequential saccades planning. Our data show robust
371 signatures of processing bottlenecks involving rate and threshold adjustments of FEF
372 movement neurons contribute to the observed processing bottlenecks. Interestingly, similar
373 adjustments of rate have been observed in FEF movement-related neurons when monkeys slow
374 their RTs to improve their accuracy (Heitz & Schall, 2012), consistent with movement related
375 activity reflecting a developing motor plan that can be adjusted by strategic requirements of
376 the task. However, in contrast, to speed/accuracy adjustments, we did find systematic increases
377 in threshold for the second saccade with shorter TSDs that together with decreases in
378 accumulation rate contribute to the lengthening of reaction times of the second saccade.
379 Interestingly, similar changes in rate for both the first and second saccade, particularly at
380 shorter TSDs, were also observed in our model of mutually inhibiting accumulators but without
381 any changes in threshold (**Fig 7**), raising the possibility that these changes may involve
382 additional processes such as adjustments in the excitability of SC neurons from the BG (Lo &
383 Wang, 2006; Wurtz & Hikosaka, 1986) that were not modelled here.
384 In contrast to movement neurons, visual neurons displayed little evidence of active
385 inhibitory control, suggesting that their activity is 'pre-bottleneck'. This is not surprising since
386 many studies have reported a separation between the visual and motor processing of FEF
387 neurons with only motor processing affecting RT in perceptually simple tasks, thus it is
388 plausible that the responses of visual neurons are not gated by inhibitory bottlenecks. This does
389 not necessarily indicate that FEF visual neurons are not subject to processing bottlenecks since
390 it is possible that in a more perceptually challenging task, manifestations of processing
391 bottlenecks would show up in the activity of visual responses. However, it can be concluded
392 that movement neurons, which are thought to be functionally downstream of visual neurons
393 (Woodman et al., 2008), are subjected to a greater degree of inhibitory control, possibly due to
394 its direct role in saccade initiation. A similar result was observed in the countermanding (Hanes,
395 Patterson, & Schall, 1998) and redirect tasks (Murthy, Ray, Shorter, Schall, & Thompson,
396 2009), where movement related neurons showed strongest evidence of inhibitory control that
397 reflect monkeys' abilities to withhold or change saccade plan. Thus, movement-related activity
398 would fall under the 'post-/peri- bottleneck' category while visually-related activity would be
399 'pre-bottleneck', at least for perceptually simple tasks.

A dichotomy between visual and movement-related activity is also consistent with these neurons being anatomically disparate, with movement neurons corresponding to layer V pyramidal neurons that project to SC and brainstem saccade generator, while the visually responsive neurons are found in both supragranular and infragranular layers (Ding & Gold, 2012; Segraves, 1992; Segraves & Goldberg, 1987; Sommer & Wurtz, 2000; Thompson, Hanes, Bichot, & Schall, 1996) showed that the caudate nucleus, and by extension the basal ganglia is involved in both perceptual decisions and saccade preparation is an oculomotor task. Input from the FEF visual neurons, that undertake perceptual decision-making, might feed into the perceptual circuit of the basal ganglia, which in turn, through the thalamo-cortical circuit can mediate a visuomotor transformation in the FEF through visuo-movement cells. Finally, as the motor command develops, local inhibitory interactions within the FEF may amplify the effect of BG resulting in more movement type neurons that instantiate the effect of the bottleneck. BG innervation of FEF within neuronal types can also allow reward value from BG reward centers to filter into the action program in the FEF, allowing for flexible constraints in neuronal dynamics and optimize motivated behavior accordingly. This speculation is also presented in (Purcell, Heitz, Cohen, & Schall, 2012; Purcell et al., 2010) wherein a gating parameter between perceptual processing by visual neurons and response preparation by movement neurons best account for both behavioral and neural data.

References

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience, 9*(1), 357-381.

Arnell, K. M., & Duncan, J. (2002). Separate and shared sources of dual-task cost in stimulus identification and response selection. *Cognitive psychology, 44*(2), 105-147.

Basu, D., & Murthy, A. (2019). Parallel programming of saccades in the macaque frontal eye field: are sequential motor plans co-activated? *Journal of neurophysiology*.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision research, 19*(9), 967-983.

Bhutani, N., Ray, S., & Murthy, A. (2012). Is saccade averaging determined by visual processing or movement planning? *Journal of neurophysiology, 108*(12), 3161-3171.

Bhutani, N., Sureshbabu, R., Farooqui, A. A., Behari, M., Goyal, V., & Murthy, A. (2013). Queuing of concurrent movement plans by basal ganglia. *Journal of Neuroscience, 33*(24), 9985-9997.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review, 113*(4), 700.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: an interactive race model of countermanding saccades. *Psychological review, 114*(2), 376.

Broadbent, D. E. (1971). Decision and stress.

Bruce, C. J., & Goldberg, M. E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. *Journal of neurophysiology, 53*(3), 603-635.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature, 487*(7405), 51-56. doi:10.1038/nature11129

Coëffé, C., & O'regan, J. K. (1987). Reducing the influence of non-target stimuli on saccade accuracy: Predictability and latency effects. *Vision research, 27*(2), 227-240.

Ding, L., & Gold, J. I. (2012). Separate, causal roles of the caudate in saccadic choice and execution in a perceptual decision task. *Neuron, 75*(5), 865-874.

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications, 7*(1), 1-15.

Findlay, J. M. (1982). Global visual processing for saccadic eye movements. *Vision research, 22*(8), 1033-1045.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience, 30*.

Goldman-Rakic, P. S., & Porrino, L. J. (1985). The primate mediodorsal (MD) nucleus and its projection to the frontal lobe. *Journal of Comparative Neurology, 242*(4), 535-560.

Gopher, D., & Navon, D. (1980). How is performance limited: Testing the notion of central capacity. *Acta psychologica, 46*(3), 161-180.

Hanes, D. P., Patterson, W. F., & Schall, J. D. (1998). Role of frontal eye fields in countermanding saccades: visual, movement, and fixation activity. *Journal of neurophysiology, 79*(2), 817-834.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science, 274*(5286), 427-430.

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron, 76*(3), 616-628.

Hikosaka, O., Takikawa, Y., & Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological reviews, 80*(3), 953-978.

Huerta, M. F., Krubitzer, L. A., & Kaas, J. H. (1986). Frontal eye field as defined by intracortical microstimulation in squirrel monkeys, owl monkeys, and macaque

monkeys: I. Subcortical connections. *Journal of Comparative Neurology, 253*(4), 415-439.

Jagadisan, U. K., & Gandhi, N. J. (2016). Disruption of fixation reveals latent sensorimotor processes in the superior colliculus. *Journal of Neuroscience, 36*(22), 6129-6140.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063): Citeseer.

Langer, T. P., & Kaneko, C. R. (1990). Brainstem afferents to the oculomotor omnipause neurons in monkey. *Journal of Comparative Neurology, 295*(3), 413-427.

Lo, C.-C., & Wang, X.-J. (2006). Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience, 9*(7), 956-963.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience, 5*(10), 793-807.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in cognitive sciences, 9*(6), 296-305.

McLeod, P. (1977). A dual task response modality effect: Support for multiprocessor models of attention. *Quarterly Journal of Experimental Psychology, 29*(4), 651-667.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain research reviews, 31*(2-3), 236-250.

Murthy, A., Ray, S., Shorter, S. M., Schall, J. D., & Thompson, K. G. (2009). Neural control of visual search by frontal eye field: effects of unexpected target displacement on visual selection and saccade preparation. *Journal of neurophysiology, 101*(5), 2485-2506.

Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive psychology, 44*(3), 193-251.

Nosofsky, R. M., & Palmeri, T. J. (1997). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception & Psychophysics, 59*(7), 1027-1048.

Parent, A., & Hazrati, L.-N. (1995a). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain research reviews, 20*(1), 91-127.

Parent, A., & Hazrati, L.-N. (1995b). Functional anatomy of the basal ganglia. II. The place of subthalamic nucleus and external pallidium in basal ganglia circuitry. *Brain research reviews, 20*(1), 128-154.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin, 116*(2), 220.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., & Schall, J. D. (2012). Response variability of frontal eye field neurons modulates with sensory input and saccade preparation but not visual search salience. *Journal of neurophysiology, 108*(10), 2737-2750.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological review, 117*(4), 1113.

Ramakrishnan, A., & Murthy, A. (2013). Brain mechanisms controlling decision making and motor planning. In *Progress in brain research* (Vol. 202, pp. 321-345): Elsevier.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of neurophysiology, 97*(2), 1756-1774.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological science, 9*(5), 347-356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review, 111*(2), 333.

Ray, S., Bhutani, N., & Murthy, A. (2012). Mutual inhibition and capacity sharing during parallel preparation of serial eye movements. *Journal of vision, 12*(3), 17-17.

Ray, S., Pouget, P., & Schall, J. D. (2009). Functional distinction between visuomovement and movement neurons in macaque frontal eye field during saccade countermanding. *Journal of neurophysiology, 102*(6), 3091-3100.

Ray, S., Schall, J. D., & Murthy, A. (2004). Programming of double-step saccade sequences: modulation by cognitive control. *Vision research, 44*(23), 2707-2718.

Ruthruff, E., Pashler, H. E., & Klaassen, A. (2001). Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement? *Psychonomic bulletin & review, 8*(1), 73-80.

Sato, T., Murthy, A., Thompson, K. G., & Schall, J. D. (2001). Search efficiency but not response interference affects visual selection in frontal eye field. *Neuron, 30*(2), 583-591.

Segraves, M. A. (1992). Activity of monkey frontal eye field neurons projecting to oculomotor regions of the pons. *Journal of neurophysiology, 68*(6), 1967-1985.

Segraves, M. A., & Goldberg, M. E. (1987). Functional properties of corticotectal neurons in the monkey's frontal eye field. *Journal of neurophysiology, 58*(6), 1387-1419.

Sendhilnathan, N., Basu, D., Goldberg, M. E., Schall, J. D., & Murthy, A. (2019). Neural correlates of goal-directed and non-goal-directed movements. *biorXiv*, 843011.

Sendhilnathan, N., Basu, D., & Murthy, A. (2017). Simultaneous analysis of the LFP and spiking activity reveals essential components of a visuomotor transformation in the frontal eye field. *Proceedings of the National Academy of Sciences, 114*(24), 6370-6375.

Sendhilnathan, N., Basu, D., & Murthy, A. (2020). Assessing within-trial and across-trial neural variability in macaque frontal eye fields and their relation to behaviour. *European Journal of Neuroscience*.

Sigman, M., & Dehaene, S. (2005). Parsing a cognitive task: a characterization of the mind's bottleneck. *PLoS biology, 3*(2), e37.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology, 53*(2), 293-315.

Sommer, M. A., & Wurtz, R. H. (2000). Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. *Journal of neurophysiology, 83*(4), 1979-2001.

Somogyi, P. (1977). A specific ‘axo-axonal’interneuron in the visual cortex of the rat. *Brain Res, 136*(2), 345-350.

Thompson, K. G., Bichot, N. P., & Schall, J. D. (1997). Dissociation of visual discrimination from saccade programming in macaque frontal eye field. *Journal of neurophysiology, 77*(2), 1046-1050.

Thompson, K. G., Hanes, D. P., Bichot, N. P., & Schall, J. D. (1996). Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *Journal of neurophysiology, 76*(6), 4040-4055.

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 3.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review, 108*(3), 550.

Viviani, P., & Swensson, R. G. (1982). Saccadic eye movements to peripherally discriminated visual targets. *Journal of Experimental Psychology: Human Perception and Performance, 8*(1), 113.

Welford, A. (1967). Single-channel operation in the brain. *Acta psychologica, 27*, 5-22.

Welford, A. T. (1952). The psychological refractory period and the timing of high-speed performance-a review and a theory. *British Journal of Psychology, 43*(1), 2.

Westheimer, G. (1954). Eye movement responses to a horizontally moving visual stimulus. *AMA archives of ophthalmology, 52*(6), 932-941.

Wheeless, L. L., Boynton, R. M., & Cohen, G. H. (1966). Eye-movement responses to step and pulse-step stimuli. *JOSA, 56*(7), 956-960.

Woodman, G. F., Kang, M.-S., Thompson, K., & Schall, J. D. (2008). The effect of visual search efficiency on response preparation: neurophysiological evidence for discrete flow. *Psychological science, 19*(2), 128-136.

Wurtz, R. H., & Hikosaka, O. (1986). Role of the basal ganglia in the initiation of saccadic eye movements. In *Progress in brain research* (Vol. 64, pp. 175-190): Elsevier.

Zambarbieri, D., Schmid, R., & Ventre, J. (1987). Saccadic eye movements to predictable visual and auditory targets. In *Eye Movements from Physiology to Cognition* (pp. 131-140): Elsevier.

Zylberberg, A., Ouellette, B., Sigman, M., & Roelfsema, P. R. (2012). Decision making during the psychological refractory period. *Current Biology, 22*(19), 1795-1799.

METHODS:

The detailed methods pertaining to this dataset has been published in a previous study (Basu & Murthy, 2019). A brief overview is given below.

Experimental Animals

Single-unit recordings were done from two adult monkeys (J, male *Macaca radiata*, and G, female *Macaca mulatta*). The animals were cared for in accordance with the animal ethics guidelines of the Committee for the Purpose of Control and Supervision of Experiments on Animals (CPCSEA), Government of India, and the Institutional Animal Ethics Committee (IAEC) of the Indian Institute of Science (IISc.).

Surgical Procedures

Each monkey underwent two surgeries: first, to implant a titanium headpost for the purpose of head-fixation during experiments, and second to make a craniotomy over the FEF and implant a recording chamber (Crist instruments, USA), whose placement was guided by an MRI scan. Training or recording sessions were conducted only after the monkeys made a complete surgical recovery.

METHOD DETAILS

Behavioral tasks:

Monkeys were trained on two oculomotor tasks: the memory-guided (MG) saccade task and the FOLLOW saccade task. Trials in the MG task began with a red fixation point (0.6° × 0.6°) which was presented in the center of a screen. After a variable fixation period(~300 ms), a gray target stimulus (1° × 1°) was presented peripherally. Post-appearance, the target disappeared after 100 ms. After target presentation, the monkeys continued to fixate during a delay period of around 1000 ms. The fixation spot was extinguished after the delay period, following which the monkeys had to make a saccade to the target location. Correct trials were reinforced with juice rewards. The delay period served to aid the classification of FEF neurons by isolating the stimulus-related (visual) and saccade-related (motor) epochs.

The FOLLOW task (**Fig. 1C**) is a modified version of the double-step task (Becker & Jürgens, 1979; Westheimer, 1954; Wheeless, Boynton, & Cohen, 1966), where single saccade no-step trials (30%) are randomly interleaved with sequential saccade step trials (70%). Trials started with fixation, following which a green saccade target (T1; 1° × 1°) was presented in one of the six possible peripheral locations (eccentricity 12º). The fixation spot was removed at target onset in no-step trials, cueing the monkeys to make a single, correct saccade to the target. In step trials the red second target (1° × 1°) appeared after T1, signaling the monkey to make ordered sequential saccades. Step trials comprised two targets, the first one being same as in the no-step trials. After a time delay (target step delay (TSD): 17 ms, 83 ms, or 150 ms), a red target (T2) was displayed. Monkeys had to make an additional second saccade from T1 to T2 to get rewarded in step trials.

Response field (RF) identification was done using the MG task. The RF center, and the two flanking positions were 'RFin' positions and the three diametrically opposite positions were considered 'RFout' positions. No-step targets and the first target of step trials could appear at any one of the six RFin and RFout locations. The second target in step trials was presented in any one of three positions diametrically opposite to the location of the first target. Based on this scheme, *RFin trials* refer to trials in which the second target or T2 was presented in the RF while T1 was outside RF. *RFout trials* are those in which T1 was inside RF and T2

was outside. Neural activity in RFin trials would mainly encode the second target or second saccade, while RFout trials would represent T1 or S1.

Recording setup and procedures:

TEMPO software (Reflective computing, St. Louis, MO, USA) was used to display and control the task paradigms. Stimuli were displayed using VIDEOSYNC software (Reflective computing) working in real time with TEMPO. Stimuli were displayed on a Sony Bravia LCD monitor (42 inches, 60 Hz refresh rate; 640 × 480 resolution) placed 57 cm from the monkeys so that 1 cm on the monitor subtended ~1°. Eye position was sampled at 240 Hz with a monocular infrared pupil tracker (ISCAN, Woburn, MA USA) that interfaced with the TEMPO software in real-time. All task parameters were saved in the TEMPO system with a sampling rate of ~1000 Hz.

Epoxy-insulated tungsten electrodes (FHC, Bowdoin, ME, USA, impedance 2 - 4 MΩ) having a diameter of 0.2 mm were used for single-unit recordings. A Cerebus data acquisition system (Blackrock Microsystems, Salt Lake City, UT, USA), which was synchronized to the TEMPO software, was used to sample and store neuronal data at 30,000 Hz. A unit was confirmed to be an FEF neuron if the activity resembled FEF cell activity types (Bruce & Goldberg, 1985). The position evoking the highest activity in the MG task was chosen as the RF of the unit. Once the neuron type and RF was characterized in the MG task (~ 100-200 trials), the FOLLOW task was initiated till sufficient data (~700 trials) was collected.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Analysis:

Neural data was sorted into individual units offline using time-amplitude algorithms provided with the Cerebus Central Suite (Blackrock Microsystems). Saccade onset and offset times were detected from the eye position data using a 30°/s velocity criterion. All analyses, after spike sorting, were done using custom-made scripts written in MATLAB (MathWorks, Natick, MA, USA).

Only correct trials, wherein saccade parameters like direction, and order, were accurate, and saccadic end points were well within the electronic windows bounding the target stimuli, were analyzed. For the FOLLOW task, trials with saccadic latency shorter than 80 ms were considered anticipatory saccades and rejected. Delayed saccades crossing the maximum permissible saccade latency limit (600 ms for no-step trials and S1 of step trials; 900 ms for second saccade of step trials) were also excluded from analysis. A minimum of 8 trials per condition was set as the criteria for performing any analysis. Units which could not be isolated or which did not show spatially selective, task-related responses were discarded. After applying the exclusion criteria, a total of 84 neurons were selected for further analysis and contributed to the results of this study.

Spike density functions (SDFs) for neural data were calculated by convolving the averaged spike train with a filter that resembled an excitatory post-synaptic potential (EPSP), having a combination of growth and decay exponential functions (Thompson et al. 1996). The time constants of the rapid growth phase (τ_g = 1 ms) and the slower decay phase (τ_d = 20 ms) were matched to values obtained from excitatory synapses and thus, were physiologically relevant.

FEF neurons were classified as visual, visuomovement, or movement by evaluating the activity during the memory-guided saccade task. Visual activity (VA) was defined as the mean firing rate in the visual epoch (90-180 ms after target onset) above the mean spontaneous firing rate measured from the baseline epoch (300-100 ms preceding target onset). Movement activity

(MA) was the mean firing rate in the movement epoch (80 ms window preceding saccade onset) above the baseline firing rate. Visual neurons significantly increased their firing rate compared to baseline in the visual epoch, post-stimulus presentation, but showed no significant activity build-up in the movement epoch. Movement neurons showed ramping up of activity prior to saccades in the movement epoch, but no stimulus-related activity. Visuomovement neurons showed heightened activity in both the visual and movement epochs.

To quantify the proportions of visual- and movement-related activity, a visuo-motor index (VMI) was calculated for each neuron (Murthy et al. 2007).

VMI was calculated as: $VMI = \frac{VA-MA}{VA+MA}$. VMI can range from +1 to -1. Visual neurons with comparatively higher VA yielded positive VMIs, whereas movement neurons with higher MA had negative VMIs. Visuomovement neurons had intermediate VMIs. In addition to the MG task, activity in the no-step trials in the FOLLOW tasks was also checked to confirm the classification, as canonical movement neurons have been shown to get transformed into visuo-motor neurons under different task contexts (Jagadisan & Gandhi, 2016).

Accumulator parameters:

Taking cue from accumulator models and previous studies (Woodman et al., 2008), four main parameters were calculated: (1) Baseline firing rate; (2) Onset of firing rate increase; (3) Threshold activity required for saccade initiation; (4) Rate of growth of activity from onset to threshold. These measures of accumulator dynamics were calculated separately for FOLLOW step trials in which the S1 went into the RF (RFout) and those in which the second saccade was towards the RF(RFin). Since correct FOLLOW step trials always had a sequence of two saccades stepping from an RF-in position to an RF-out position or vice versa, the activity in the RFin trials was a mix of RFout and RFin activities. To specifically analyze activity that contributed only to the second saccades made into the RF, the mean RF-out activity of no-step trials was used as a reference and subtracted from the mean activity in RFin step-trials. For single neurons, the parameter calculations were made from non-normalized, differential activity for RFin trials.

For FOLLOW no-step trials, trials in which the saccade was into the response field were used for accumulator parameter calculation. Trials in each session were grouped into fast, medium, or slow RT trials based on the average RT of that session, i.e. trials with RT less than 30ms below mean RT were considered as fast trials, those with RT more than 30 ms above mean RT were slow trials, and trials around the mean RT (± 10 ms) were medium RT trials. Accumulator parameters were then calculated for the three RT groups.

Baseline activity was measured as the average of the differential activity in the RFin condition (and raw activity in the RFout/no-step condition) in the 100 ms before the appearance of the first FOLLOW target. Onset was defined as the time point when FEF activity first exceeded 2 SDs above baseline, provided that the differential activity ultimately reached 4 SDs and was maintained above 2 SDs for at least 50 ms for the second saccade plan and 20 ms for the S1 plan. Threshold activation was the average firing rate in the RF-in condition in the interval from 10 to 20 ms before saccade initiation (Hanes & Schall, 1996). Rate of activity growth was measured by subtracting the threshold-activity level from the onset-activity level and dividing by the time interval between onset and threshold. This measure was robust against fluctuations in the rise-profile. To better understand non-linear rise profiles, the rate was also measured by piecewise regression fits using a sliding window of width 40 ms (for RFin trials) from onset to threshold and calculating the slopes and intercepts at each point. For population analyses, difference SDFs of each session were normalized to the peak average activity in the TSD = 17 ms group and for each saccade plan i.e. activity related to second saccade plan was normalized with respect to TSD 17 activity for second saccades going into the RF and vice

versa for the S1 plan. In case of no-step trials, the SDFs were normalized to the peak activity of the fast trials in each session.

Principal Components Analysis (PCA):

For analyses based on dimensionality reduction, we performed two steps of data preprocessing before further analyses. First, we 'soft normalized' the neural responses (r) for each neuron, i, by dividing the neural activity for each neuron (r_i) by its range ($|r_i| = r_i/(range(r_i))$) (Churchland et al., 2012). Soft normalization preserves the structure of inter-neuronal variation while normalizing the population response so that neurons with strong responses could be reduced to approximately unity range, but neurons with weak responses could be reduced to less than unity range. Second, we mean-centered the responses of each neuron by subtracting the mean activity of a given neuron across all conditions ($\overline{r_i}$) from the neural response ($|r_i| - \overline{r_i}$).

To identify the signals that best represent the population activity of neurons, we performed principle component analysis (PCA), a common unsupervised learning algorithm, on the data. To do this, we constructed two matrices P_1 and P_2 of size $t \times n$ where t is the time and n is the number of neurons with population response for the first and the second saccade plans respectively. We applied PCA to P_1 and P_2 yielding W_1 and W_2 respectively, which are $n \times k$ matrix each, of principal components.

We used a metric to index the degree to which the population response occupied different neural dimensions on trials with different TSDs. To compute this 'subspace overlap' between the first and the second saccade plans for each TSD, we first defined the variance captured as $V(P, W) = 1 - \frac{\|P - PWW^T\|}{\|P\|}$, where the operator ||X|| means the Frobenius norm of the matrix X. Then, the subspace overlap was given by: $\frac{V(P_2W_1)}{V(P_2W_2)}$.

Subspace overlap should be equal to one if the population responses occupied the same dimensions (i.e., are spanned by the same PCs) on both the saccade plans. And the subspace overlap should be equal to zero if the population responses occupied mutually orthogonal dimensions on both the saccade plans.

Neural Simulations:

We simulated 40 motor neurons with 900 trials (with three types of TSD trials) using a firing rate model to approximately match the statistical power of our experimental dataset. We defined the spatial properties of each neuron (tuning curve) through a cosine function centered on one of the 8 positions which was randomly chosen and called it the neuron's RF:

$$tuning(\theta) = r\cos(w\theta + \phi)$$

where r is the peak firing rate, w defines the width of the tuning curve, θ is a set of 8 target positions and ϕ is the displacement (**Fig S4B**). We then defined the temporal properties of the neurons (**Fig S4C**) using a skewed Gaussian distribution response kernel:

$$kernel\ (x) = 2\frac{1}{\sqrt{2\pi}}\, e^{\frac{-(x-m)^2}{2c}} \int_{-\infty}^{\alpha x} \frac{1}{\sqrt{2\pi}}\, e^{\frac{-t^2}{2c}}\, dt$$

$$= \frac{1}{\sqrt{2\pi}}\, e^{\frac{-x^2}{2c}} \left[1 + erf\left(\frac{\alpha x}{\sqrt{2}}\right)\right]$$

where $c = 8000$ controls the full width at half maximum of the distribution, $erf(x)$ is the error function and α is the shape parameter that controls the shape of the distribution. To simulate a noisy accumulator (**Fig S4H**), we sampled $m \sim \mathcal{N}(0, 40)$ and $\alpha \sim -|\mathcal{N}(0, 0.01)|$. Note that similar results can be obtained using a response kernel resembling a Poisson post-synaptic potential function:

$$kernerl\ (x) = \left[1 - e^{\frac{-x}{\tau_g}}\right] \times \left[e^{\frac{-x}{\tau_d}}\right]$$

where τ_g controls the rate of increase and τ_d controls the rate of decay.

We further injected a small noise to the system by convolving the response kernel with Gaussian noise of $\mu = 0$ and $\sigma^2 = 1$. For each neuron, we multiplied the tuning curve and the convolved kernel to get the spatiotemporal firing rate for that neuron (**Fig S4D**). Saccade onsets were taken as the time when the normalized activity reached a fixed threshold of 1 unit.

For the simulated data without mutual inhibition, we simply followed the above steps for the plans for saccade 1 and saccade 2 and linearly added them on each trial. For the simulated data with mutual inhibition, we followed the same steps (for both plans 1 and 2) as above until the activity for plan 1 reached y% of its peak response. This 'y' is given by an inhibition function (**Fig S4G**) which was constructed by dividing the neural response by the response in no-step trials. To simulate the inhibitory effect of plan 1 on plan 2, we temporally shifted the response for plan 2, to until after the time of first saccade onset (t_{s1}), after plan 1 reached y% of its peak response and then interpolated the data in between using a cubic spline function.

$$r_2(t) = \begin{cases} r_2(t), & \text{if } t < T \\ r_2(t + (t_{s1} - T)), & \text{if } t \geq t_{s1} \end{cases}$$

where $T = arg\ \left(y \times \max\left(r_1(t)\right)\right)$

To simulate the effect of plan 2 on plan 1, we first normalized plan 2's response from 0 to 1, then multiplied it by the same inhibition factor, y and then subtracted it from plan 1's response.

$$r_1(t) = r_1(t) - y\left(\frac{r_2(t) - \min\left(r_2(t)\right)}{\max\left(r_2(t)\right) - \min\left(r_2(t)\right)}\right)$$

In this way, plans 1 and 2 have the highest mutual inhibition effect on each other for the shortest TSD and the least effect for the longest TSD. We estimated all the above hyper-parameters such that the simulated data closely resembled the experimental data. After simulating the data, we followed the same data pre-processing step before dimensionality reduction similar to the experimental data.

Statistical testing:

A two-sided Wilcoxon signed-rank test to analyze a single sample set of data. For group comparisons, the non-parametric Kruskal-Wallis test was used. Trials were considered to be independent observations as the TSDs for each trial were chosen randomly. All the results are presented as mean (± standard error of mean, SEM) and all tests are performed at a significance level of $\alpha = 0.05$ unless otherwise mentioned.

DATA AND CODE AVAILABILITY:

All data is available in the main text or the supplementary materials. Raw data and codes are available upon reasonable request.

Acknowledgments:
We thank S. Sengupta for helping with behavioral training and Dr. A. Gopal P.A. for helping with data collection.

Funding:
This work was supported by a D.B.T.-I.I.Sc (Department of Biotechnology, Government of India – Indian Institute of Science) partnership grant given to A.M. D.B was supported by a graduate fellowship from the Ministry of Human Resource Development (MHRD), Government of India, through the Indian Institute of Science.

Author Contributions:
Conceptualization, A.M.
Methodology, D.B. & N.S.
Investigation, D.B.
Formal analysis, D.B. & N.S.
Writing - Original Draft, D.B., N.S. & A.M.
Writing - Review & Editing, D.B., N.S. & A.M.
Visualization, D.B. & N.S.
Supervision, A.M.
Project administration, A.M.
Funding acquisition, A.M.

Competing interests: Authors declare no competing interests.

Disclosure: The authors declare no competing financial interests

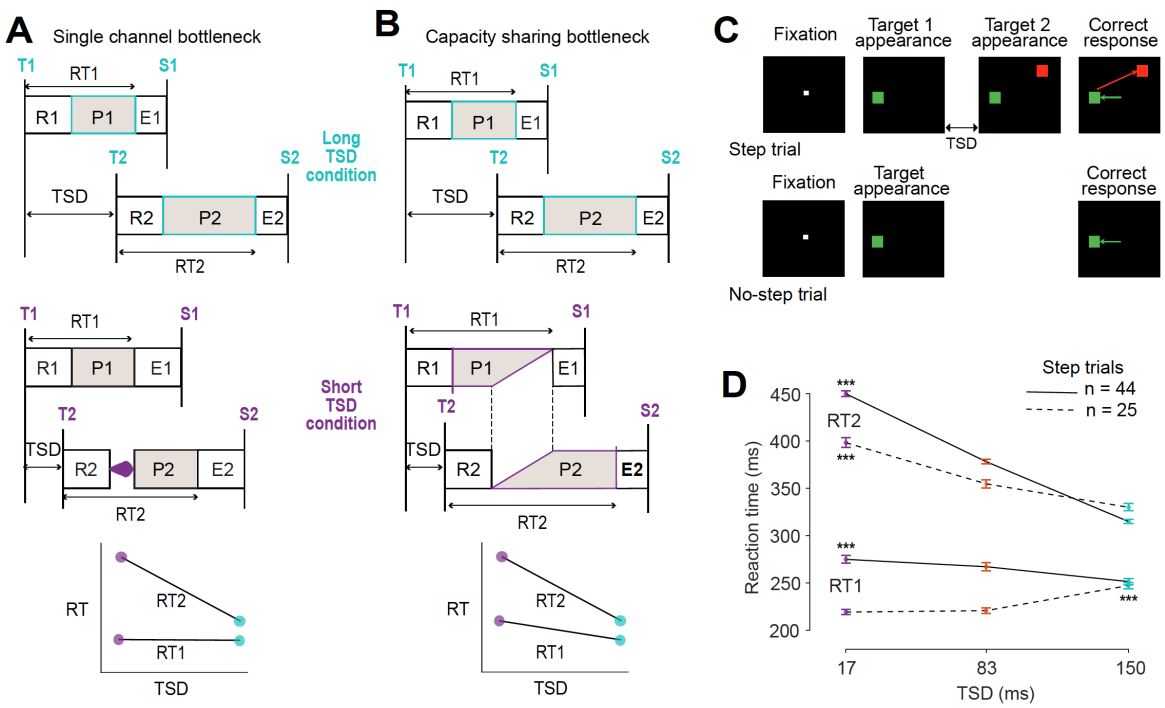

Figure 1. Framework of processing bottlenecks during the planning of sequential saccades

A. Single channel bottleneck. In a simplistic framework, the processes leading to the culmination of a saccade are assumed to consist of three stages: Stimulus Registration (R), Central Planning (P), and Saccade execution (E). T1 & T2 refer to the first and second target onset and S1 and S2 refer to the first and second saccade end, respectively. Reaction time, RT, consists of the R & P stages. According to the single-channel bottleneck model, the stages of the first saccade plan (T1-S1) proceed to completion unabated. When the T2 appears after a variable target step delay TSD, the S2 plan is initiated, however the central planning stage, P2, can only proceed after P1 is complete. Thus, as depicted in the schematic, as the TSD decreases from long (top panel) to short (middle panel), the imposed bottleneck postpones P2 progressively, leading to the increase of RT2 with TSD.
Bottom panel: Postponement of the P2 stage brings about the RT2-TSD relation shown in the schematic plot. Increase of RT2 with decrease in TSD is the primary behavioral evidence of processing bottlenecks. Concomitant RT1 increase is absent in the single-channel bottleneck framework.

B. Capacity sharing bottleneck. In this alternate framework, the P2 stage can continue in parallel with P1, and thus the central planning stages are not discretized. Under long TSD conditions (top panel), there is no overlap between the central stages of the two saccade plans and thus, both plans get 100% of the brain's limited processing capacity. When the TSD is reduced (middle panel), there is a greater degree of overlap between the saccade plans. P1 and P2 can both proceed in parallel, sharing the processing capacity. Since both saccade plans can only access a part of the entire capacity, the central planning stages become elongated compared to the high TSD condition, and both RT1 & RT2 increase. The lengthening of RT2 is much higher (not depicted in the illustration for simplicity) as the S1 plan is already underway, having taken the major share of the limited capacity and thus the S2 plan gets a much smaller allocation and takes longer time to reach completion (bottom manel).

C. FOLLOW task. Examples of a step trial (top panel) and a no-step trial (bottom panel) of the FOLLOW task are shown. Both trial types start with central fixation, followed by the presentation of the first green target. In step trials, a second red target appears after a variable TSD. The monkeys had to make saccades to the targets in order of their appearance to earn reward.

D. Task behavior. The S2 RT increased significantly with decrease in TSD, displaying the classical behavioral evidence of processing bottlenecks. The S1 RT increased slightly but significantly, with decrease in TSD for a subset of sessions (n = 44).

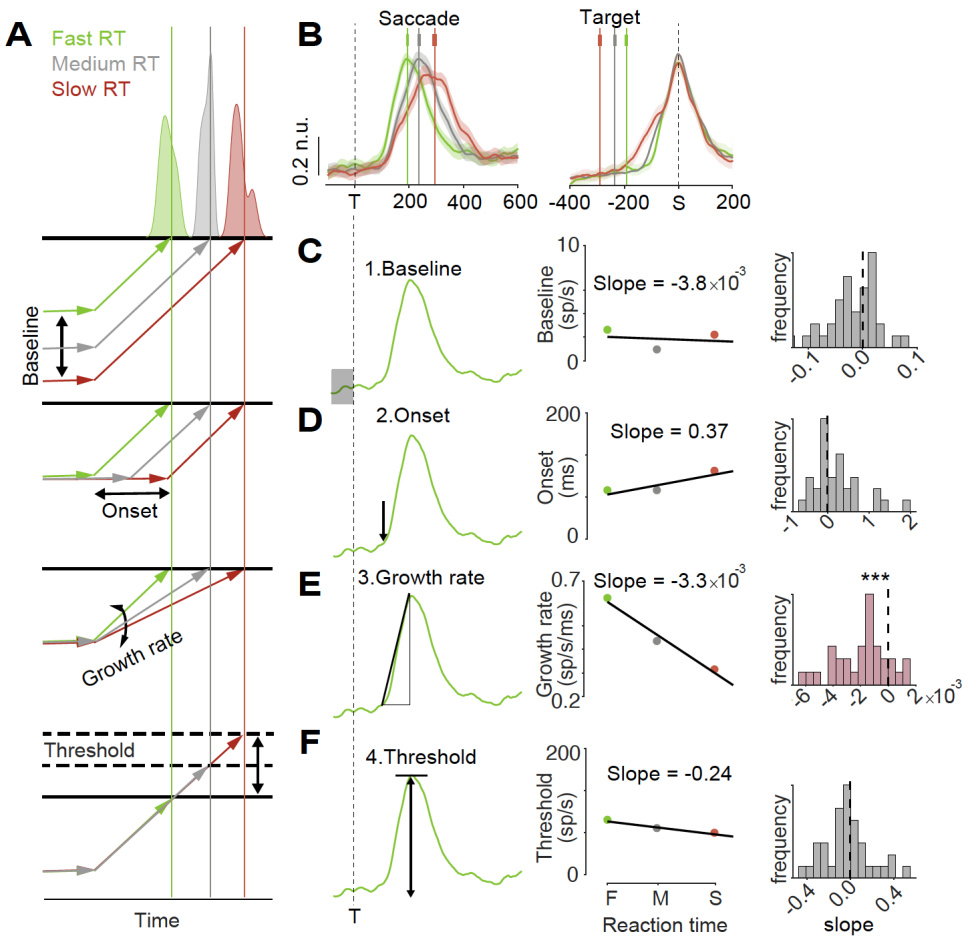

Figure 2. Population activity of movement neurons when the saccade went into RF in no-step trials

A. Schematic of possible adjustments of neural activity related to processing bottlenecks. The lengthening of the RT may be due to four possible adjustments according to accumulator dynamics. Each schematic shows a noise-free, simplistic accumulation process, starting from the baseline and reaching up to the threshold for saccade initiation. Representative RT distributions are plotted above each of the schematics. Lowering of baseline activity, delaying the onset of activity, slowing down the rate of growth, and increasing the threshold level can either singly or in combination, bring about increased reaction times.

B. Left: FEF movement neuron population activity encoding for different RTs (short: green, medium: gray, long: red), aligned to target onset (vertical broken line) . Mean saccade onset times for all the RT conditions are shown as vertical colored lines with s.e.m error bars. Right: same as left but activity aligned to saccade onset (vertical broken line).

C. Illustration of the measurement of accumulator parameter: baseline. Left: Schematic of a spike density function of a representative neuron aligned to target onset showing the baseline value. Middle: The baseline was calculated for short, medium and long RTs for a representative neuron and the slope of the best fit was calculated. Right: Histogram of distribution of slopes measured this way for all the neurons was compared with zero (vertical broken line).

D. Same as **Fig 2C** but for measurement of onset.

E. Same as **Fig 2C** but for measurement of growth rate.

F. Same as **Fig 2C** but for measurement of threshold.

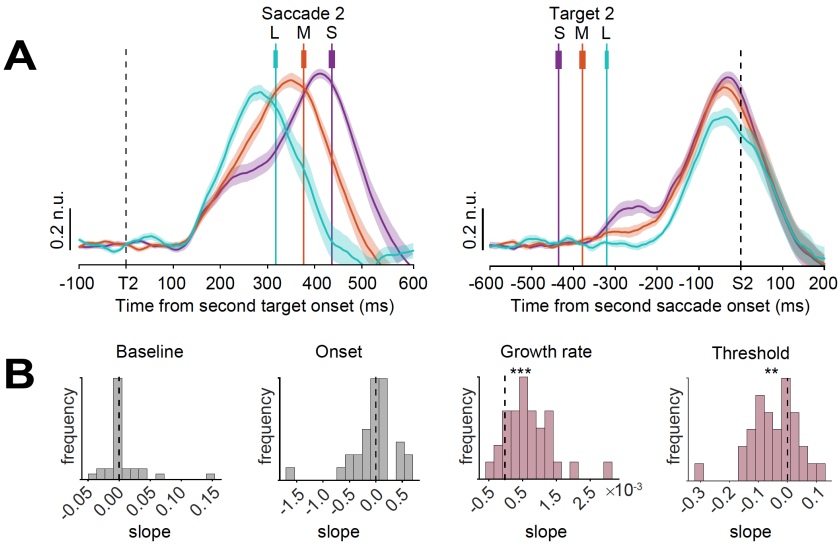

Figure 3. Population activity of movement neurons when the second saccade went into the RF in step trials

A. Left: FEF movement neuron population activity encoding for different TSD conditions (short: purple, medium: red, long: turquoise), aligned to T2 onset (vertical broken line). Mean saccade onset times for the TSD conditions are shown as vertical colored lines with s.e.m error bars. Right: same as left but activity aligned to S2 onset (vertical broken line).

B. Population histogram of slopes of each measure of accumulator dynamics (baseline, onset, growth rate and threshold) as a function of TSD for FEF movement neurons. Asterisks denote cases where the distribution of movement neuron slopes was significantly different from zero (Wilcoxon signed-rank test, ***$p < .001$, **$p < .01$).

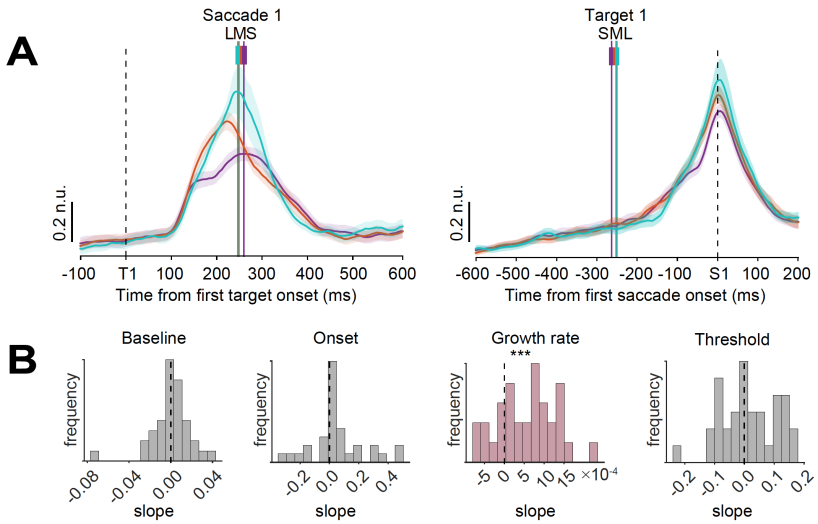

Figure 4. Population activity of movement neurons when the first saccade went into the RF in step trials

A. Left: FEF movement neuron population activity encoding for different TSD conditions (short: purple, medium: red, long: turquoise), aligned to T1 onset (vertical broken line) . Mean saccade onset times for the TSD conditions are shown as vertical colored lines with s.e.m error bars. Right: same as left but activity aligned to S1 onset (vertical broken line).

B. Population histogram of slopes of each measure of neural dynamics (baseline, onset, growth rate and threshold) as a function of TSD for FEF movement neurons. Asterisks denote cases where the distribution of movement neuron slopes was significantly different from zero (Wilcoxon signed-rank test, ***$p < .001$).

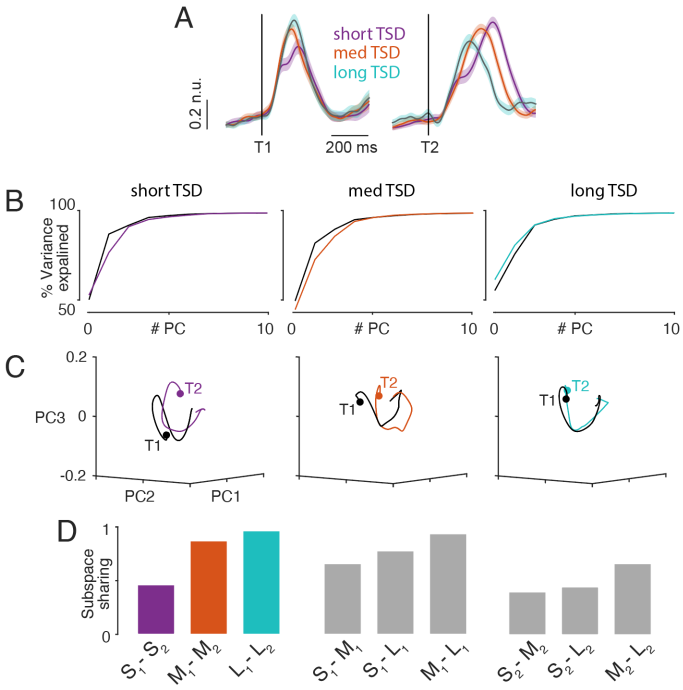

Figure 5: Extent of subspace sharing explains processing bottlenecks during the planning of sequential saccades

A. Normalized mean population neural responses aligned to T1 and T2 for short, medium and long TSD trials (n.u. = normalized unit).

B. Cumulative percent variance explained by the first 10 PCs for T1 (black) and T2 (color) related responses for short (left), medium (center) and long (right) TSD trials.

C. First three principle components (PC) plotted against each other for T1 (black) and T2 (color) related responses for short (left), medium (center) and long (right) TSD trials Filled circle markers indicate the starts of the respective trajectories.

D. Subspace overlap between a pair of conditions. S, M and L indicate short, medium and long TSDs, and 1 and 2 indicate the saccade plan number.

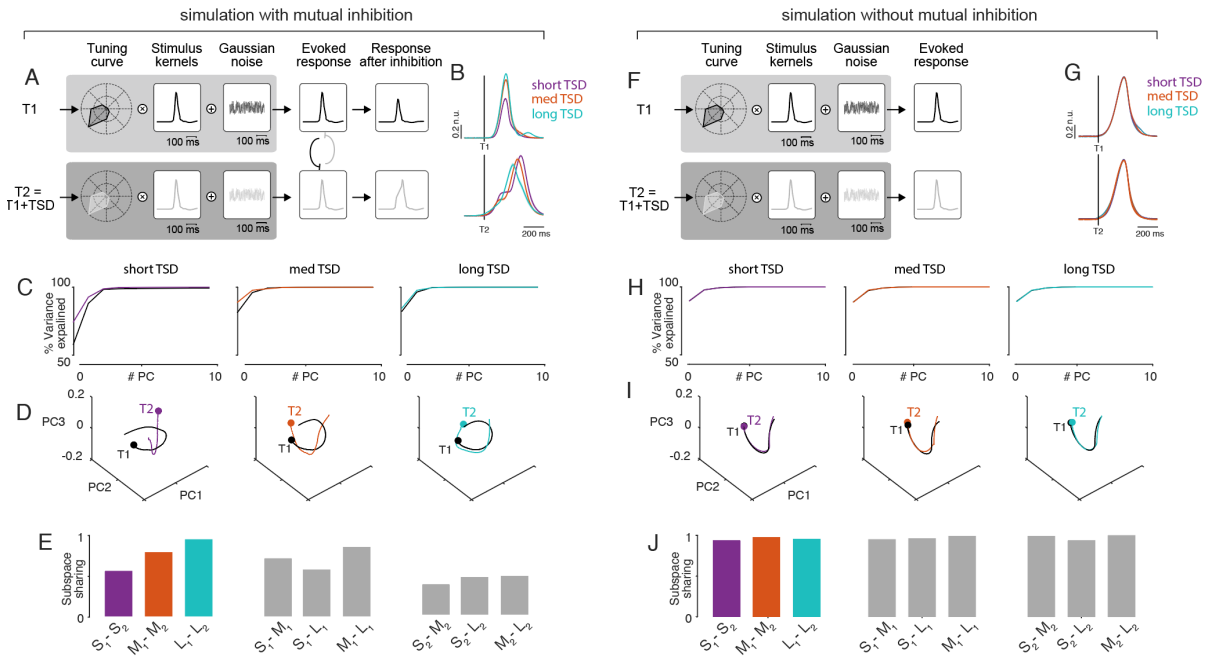

Figure 6: Simulations with mutual inhibition capture the empirical data's population dynamics

- **A.** Schematic of the simulation with mutual inhibition. Top row: simulated neural activity for the S1 plan and bottom row: simulated neural activity for the S2 plan (see **Fig S3A-G** and methods; n.u. = normalized unit).
- **B.** Normalized mean population neural responses, for data simulated with mutual inhibition, aligned to T1 and T2 for short, medium and long TSD trials.
- **C.** Cumulative percent variance explained by the first 10 PCs for T1 (black) and T2 (color) related responses for short (left), medium (center) and long (right) TSD trials.
- **D.** First three PCs plotted against each other for T1 (black) and T2 (color) related responses for short (left), medium (center) and long (right) TSD trials. Filled circle markers indicate the starts of the respective trajectories.
- **E.** Subspace overlap between a pair of conditions. S, M and L indicate short, medium and long TSDs, and 1 and 2 indicate the plan number.
- **F.** Same as **Fig 6A,** but for simulation without mutual inhibition (see methods).
- **G.** Same as **Fig 6B**, but for simulated data without mutual inhibition.
- **H.** Same as **Fig 6C**, but for simulated data without mutual inhibition.
- **I.** Same as **Fig 6D**, but for simulated data without mutual inhibition.
- **J.** Same as **Fig 6E**, but for simulated data without mutual inhibition.

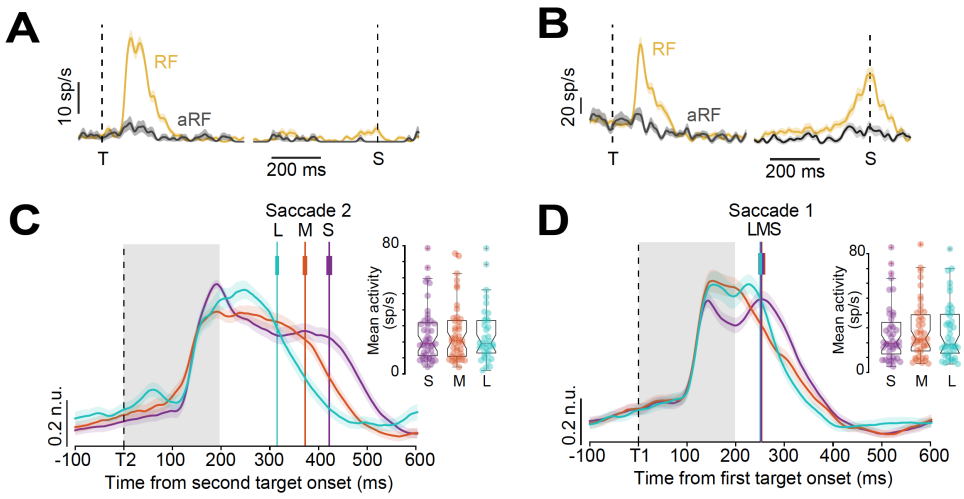

Figure 7. Processing bottlenecks in FEF V and VM neurons

A. A representative FEF visual neuron aligned to target onset (T) and saccade onset (S) in a memory-guided task for saccades into the RF (yellow) and saccades out of RF (into aRF; black).

B. A representative FEF vismov neuron aligned to target onset (T) and saccade onset (S) in a memory-guided task for saccades into the RF (yellow) and saccades out of RF (into aRF; black).

C. FEF visual and vismov neuron population activity encoding for different TSD conditions (short: purple, medium: red, long: turquoise), aligned to T2 onset (vertical broken line). Mean saccade onset times for the TSD conditions are shown as vertical colored lines with s.e.m error bars. Inset: Kruskal–Wallis box plots for short, medium and long TSDs.

D. Same as **Fig 7C** but for neural activity aligned to T1.

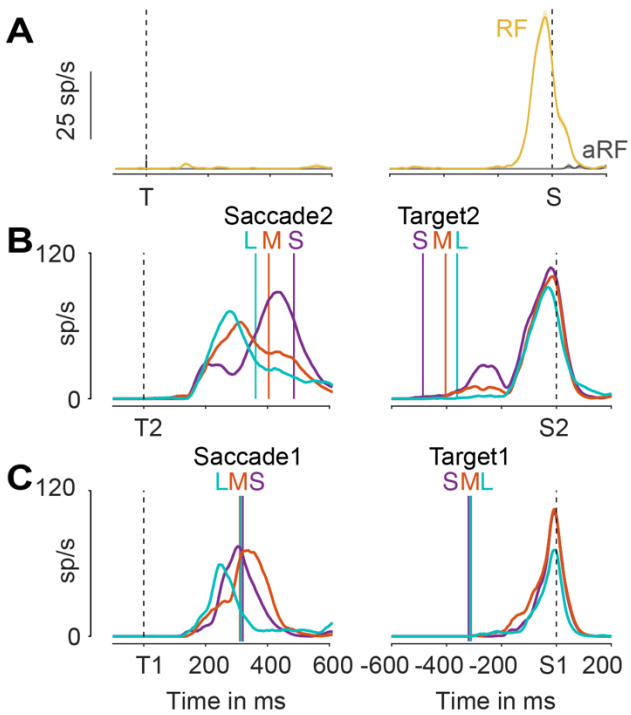

Figure S1 (related to Fig 3 & 4): Single neuron example for processing bottlenecks

A. A representative FEF movement neuron aligned to target onset (T) and saccade onset (S) in a memory guided task for saccades into the RF (yellow) and saccades out of RF (into aRF; gray).
B. Left: Neural activity, of the same representative neuron from **Fig S1A,** encoding for different TSD conditions (short: purple, medium: red, long: turquoise), aligned to T2 onset (vertical broken line). Right: same as left but activity aligned to S2 onset (vertical broken line).
C. Same as Fig S1B but for neural activity aligned to T1 (left) and S1 (right).

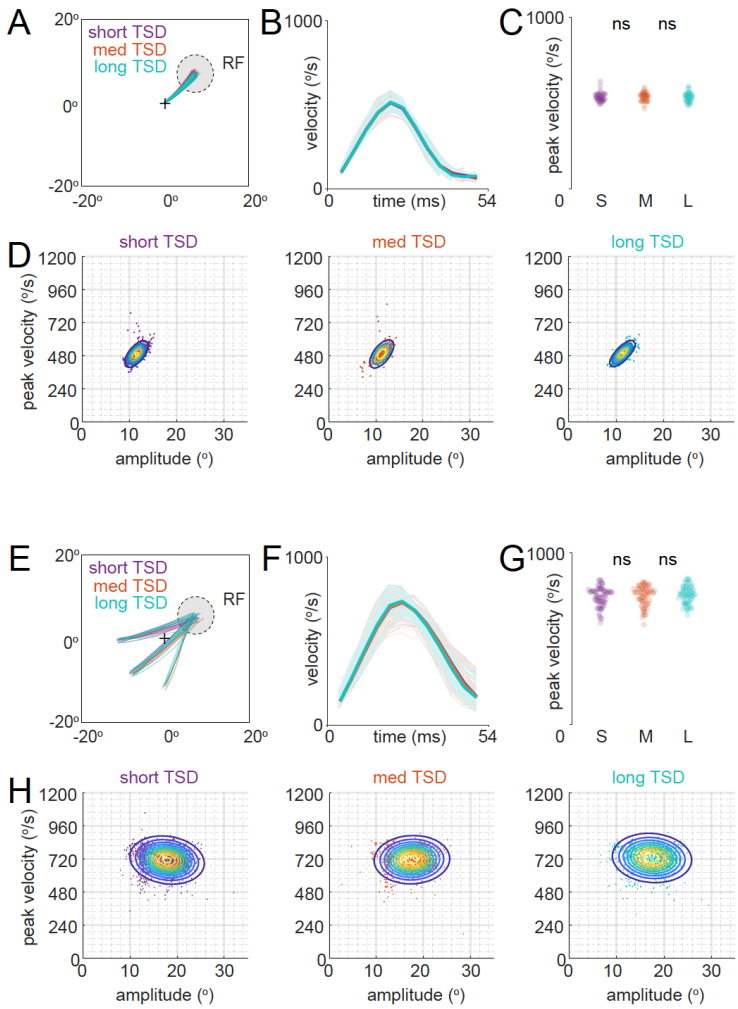

Figure S2 (related to Fig 5): Saccade kinematics for the S1 and S2 with different TSDs

A. Saccade 1 trajectories, towards RF (filled, broken circle), during short, medium and long TSD trials from a representative session. The central cross denotes the fixation point.

B. Mean saccade 1 velocity profiles for short, medium and long TSD trials (thick lines) superimposed on saccade 1 velocity profiles from individual trials (thin lines).

C. Peak saccade 1 velocities from each session for short, medium and long TSD trials. P values: short-medium: 0.81, ranksum test; medium-long: 0.60, ranksum test.

D. Saccade 1 main sequence for short (left), medium (center) and long (right) TSD trials. The overlaid contours represent the density of data.

E. Saccade 2 trajectories, towards RF (filled, broken circle), during short, medium and long TSD trials from the same representative session as **Fig S2A**.

F. Same as **Fig S2B**, but for saccade 2.

G. Same as **Fig S2C**, but for saccade 2. P values: short-medium: 0.92, ranksum test; medium-long: 0.80, ranksum test.

H. Same as **Fig S2D**, but for saccade 2.

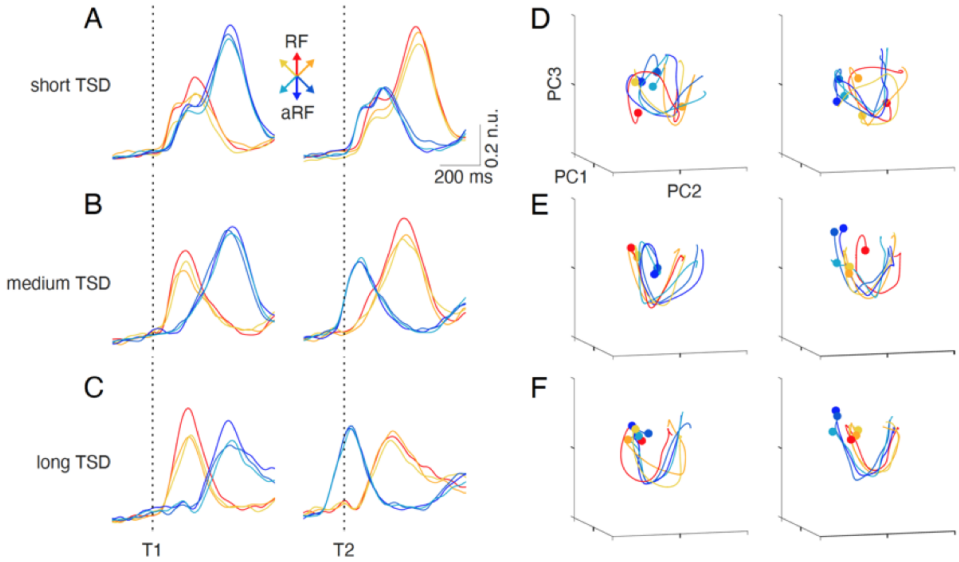

Figure S3 (related to Fig 5): Population dynamics for all target positions

A. Soft-normalized mean population neural responses towards all six target positions (3 towards RF, shown in warm colors and 3 out of RF, shown in cold colors; see inset for colors), aligned to T1 (left) and T2 (right) for short TSD trials (n.u. = normalized unit).

B. Same as **Fig S3A**, but for medium TSD trials.

C. Same as **Fig S3A**, but for long TSD trials.

D. First three PCs for each target position shown in **Fig S3A**, plotted against each other for T1 (left) and T2 (right) related responses for short TSD trials. Filled circle markers indicate the starts of the respective trajectories.

E. Same as **Fig S3D**, but for medium TSD trials shown in **Fig S3B**.

F. Same as **Fig S3D**, but for long TSD trials shown in **Fig S3C**.

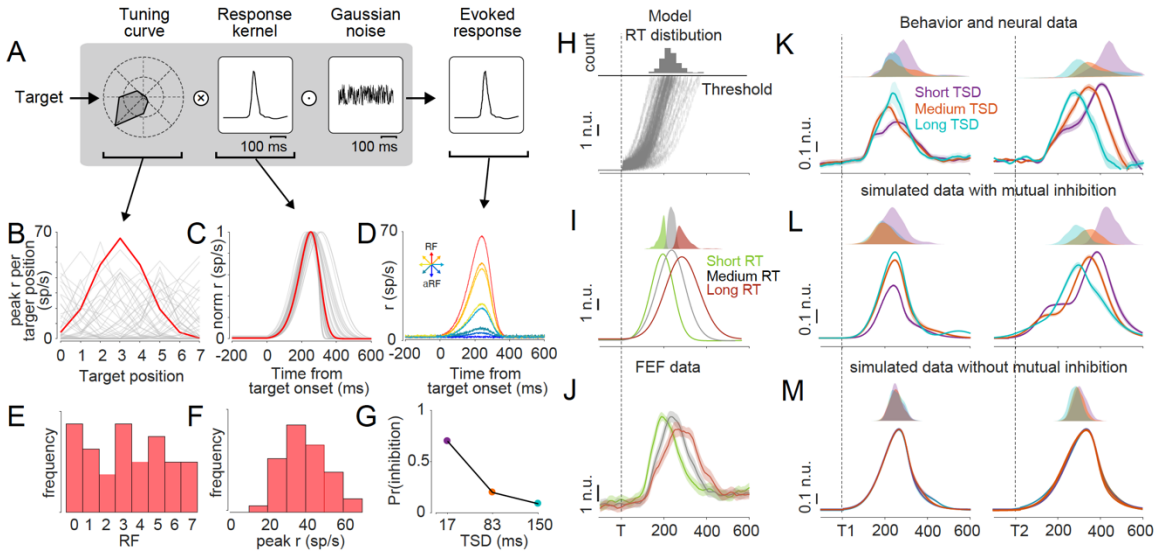

Figure S4 (related to Fig 6): Data simulation process and RT profiles

A. Schematic illustration of the simulation.

B. Tuning curves (peak firing rates as a function of target positions) for all the simulated neurons. One representative neuron's tuning curve is highlighted in purple.

C. Response kernels (activity as a function of time) for all the simulated neurons. Same representative neuron's stimulus kernel is highlighted in purple.

D. Multiplication of tuning curves and stimulus kernels and addition of noise results in the simulated neuron's activity for each of the 8 target positions. Activity towards the 8 target positions is shown for the same representative neuron as above.

E. Distribution of RF positions (argmax(tuning curve)) of the simulated neurons shows a uniform spread of RFs across the simulated neurons.

F. Distribution of peak firing rate (max(tuning curve)) of the simulated neurons shows a unimodal distribution with a mean ~ 37 sp/s.

G. Probability of mutual inhibition as a function of TSD

H. Bottom: Example simulation for one session, for RF position, following the simulation pipeline shown in A. Top: RT distribution obtained from the simulation.

I. Top: The same RT distribution from H, but uniformly divided into fast, medium and slow RTs. Bottom: Average neural activity from the simulation in H, divided into fast, medium and slow RT conditions as explained above.

J. Same FEF neural data from **Fig 2B** (divided into fast, medium and slow RT).

K. Top: RT distributions of two monkeys for short, medium and long TSD trials for plan 1 (left) and plan 2 (right). Bottom: left: Same as **Fig 4A left;** right: Same as **Fig 3A left.**

L. Top: RT distributions from the simulations with mutual inhibition in the same format as above. Bottom: Same as **Fig 6B.**

M. Top: RT distributions from the simulations without mutual inhibition in the same format as above. Bottom: Same as **Fig 6G.**

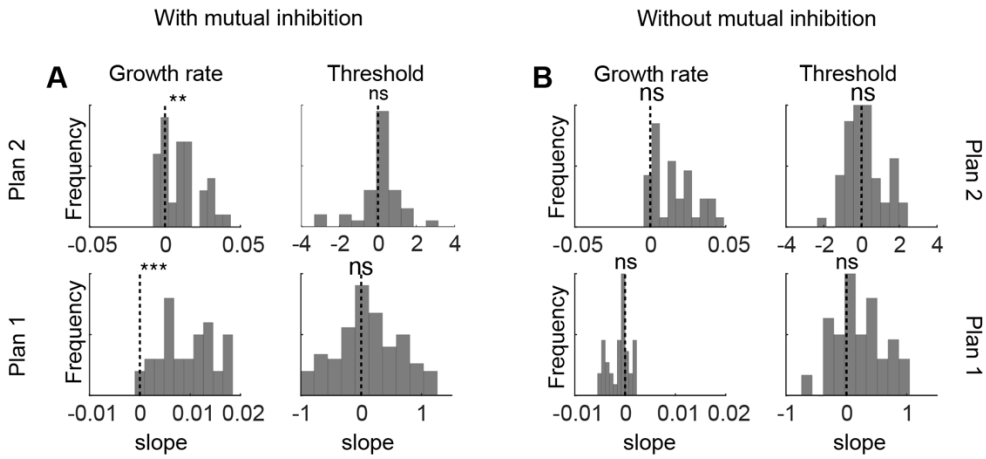

Figure S5 (related to Fig 6): Model with mutual inhibition replicates neural data

A. **.** Top: Population histogram of slopes of measures of accumulator dynamics (growth rate and threshold) during plan 2 as a function of TSD for simulated neurons (with mutual inhibition). Asterisks denote cases where the distribution of movement neuron slopes was significantly different from zero (Wilcoxon signed-rank test, ***$p < .001$, **$p < .01$). Bottom: Same as top panel but for plan 1.

B. Same as **S5A** but for model without mutual inhibition