PSYCHOLOGICAL SCIENCE

A unified model of the task-evoked pupil response

Charlie S. Burlingham[1]*†, Saghar Mirbagheri[2]*†, David J. Heeger[1,3]*

The pupil dilates and reconstricts following task events. It is popular to model this task-evoked pupil response as a linear transformation of event-locked impulses, whose amplitudes are used as estimates of arousal. We show that this model is incorrect and propose an alternative model based on the physiological finding that a common neural input drives saccades and pupil size. The estimates of arousal from our model agreed with key predictions: Arousal scaled with task difficulty and behavioral performance but was invariant to small differences in trial duration. Moreover, the model offers a unified explanation for a wide range of phenomena: entrainment of pupil size and saccades to task timing, modulation of pupil response amplitude and noise with task difficulty, reaction time–dependent modulation of pupil response timing and amplitude, a constrictory pupil response time-locked to saccades, and task-dependent distortion of this saccade-locked pupil response.

INTRODUCTION

Although the pupil responds most strongly to changes in luminance (*1*, *2*) and accommodation (*1*, *3*), it fluctuates in size even in their absence (*4*). During fixation and in the absence of a task, pupil size changes constantly and in a seemingly random way (fig. S1A) (*4*, *5*). During task performance, pupil size entrains to trial timing, dilating and then constricting sluggishly following trial onsets (fig. S1C) (*6*, *7*). This so-called "task-evoked" pupil response is modulated in amplitude by task demands, behavioral performance, and surprise (*7*–*9*). Pupil responses covary with sweating and cardiac activity (*10*); measures of peripheral autonomic activity; and spiking activity in the locus coeruleus (LC) (*8*, *11*, *12*), basal forebrain (*11*), and dorsal raphé (*13*), which are sources of cortical norepinephrine (NE), acetylcholine, and serotonin, respectively. On the basis of these observations, many studies have used pupil size as a noninvasive measure of arousal level (*2*, *14*–*16*).

The task-evoked pupil response is commonly used to estimate arousal by assuming that pupil size is a linear transformation of task events (e.g., stimulus onset, button press, and feedback). Specifically, existing models (Fig. 1A) assume that pupil size is the output of a dilatory low-pass filter, which operates on a series of impulses aligned with task events (*17*–*22*). Arousal level, estimated as the amplitude of these inputs, is estimated using linear regression. We will refer to this approach as the "consensus model," acknowledging that there are important predecessors in the literature (*23*, *24*). Although the consensus model is widely used, its central assumption that pupil size is a linear transformation of task events has not been tested.

We tested the consensus model with a visual orientation-discrimination task in humans. The task was either easy or hard, to manipulate arousal, and had a short or long wait time between trials (2 or 4 s), which was expected to have little influence on arousal. Estimates of arousal from the consensus model failed to generalize between 2- and 4-s trials and were highly variable, demonstrating that pupil size is not a linear transformation of task events. We propose an alternative model linking the task-evoked pupil response and arousal, which leverages the "common drive hypothesis"—i.e., that pupil size and saccades are driven by a common neural input—to constrain estimation of arousal (*12*, *16*, *25*–*27*).

There is substantial support for the common drive hypothesis. A transient pupil constriction and redilation locked to saccades was reported in humans as early as the 1960s (*1*, *28*–*30*). Microstimulation of superior colliculus (SC), a subcortical center for saccade generation, evokes a transient pupil response for electrical stimulation current levels above or below that required to evoke a saccade (*12*, *25*, *27*). Furthermore, converging anatomical (*2*, *26*, *31*), physiological (*1*, *12*, *26*, *32*), and pharmacological evidence (*4*, *33*–*36*) supports predominantly parasympathetic control of the task-evoked pupil response via the preganglionic Edinger-Westphal (EW) nucleus, which receives direct projections from the SC (*26*). Involuntary fixational saccades (including microsaccades in our definition), like pupil responses, fluctuate with engagement—entraining to task timing. Indeed, pupil size and (micro)saccades are correlated in a wide array of tasks (*6*, *7*, *9*, *37*, *38*). Specifically, saccade occurrence is suppressed nearly completely at trial onset, resumes quickly thereafter, and then is more slowly suppressed again in anticipation of the next trial (*38*–*40*). The magnitude of this oculomotor suppression has been linked to temporal attention (*38*, *41*), temporal expectation (*41*), task difficulty (*40*), and perceptual detection (*42*), task variables that have also been linked to pupil responses (*7*, *9*, *17*, *22*, *24*).

Our linking model has two channels (Fig. 1B). The first relates the common input to saccades, and the second relates the input to pupil size. In the first channel, a saccade is generated whenever a noisy generator function crosses a fixed threshold. In the second channel, the same generator function is amplified by a gain and low-pass–filtered, yielding pupil size. The gain and generator function may vary systematically from trial to trial and from task to task. We offer an algorithm, based on this model, that estimates the gain and generator function from measurements of pupil size and saccade rate.

Our results strongly support the hypothesis that pupil size and saccades are driven by a common input (*16*, *26*, *27*) and extends that hypothesis to account for the influence of arousal on the task-evoked pupil response. Saccade timing, transformed with our model, accurately predicted the task-evoked pupil response. Gain, our model's estimate of arousal, scaled with task difficulty and behavioral performance and generalized well between trials with different

[1] Department of Psychology, New York University, New York, NY 10003, USA. [2] Graduate Program in Neuroscience, University of Washington, Seattle, WA 98195, USA. [3] Center for Neural Science, New York University, New York, NY 10003, USA.
*Corresponding author. Email: charlie.burlingham@nyu.edu (C.S.B.); sagharm@uw.edu (S.M.); david.heeger@nyu.edu (D.J.H.)
†These authors contributed equally to this work.

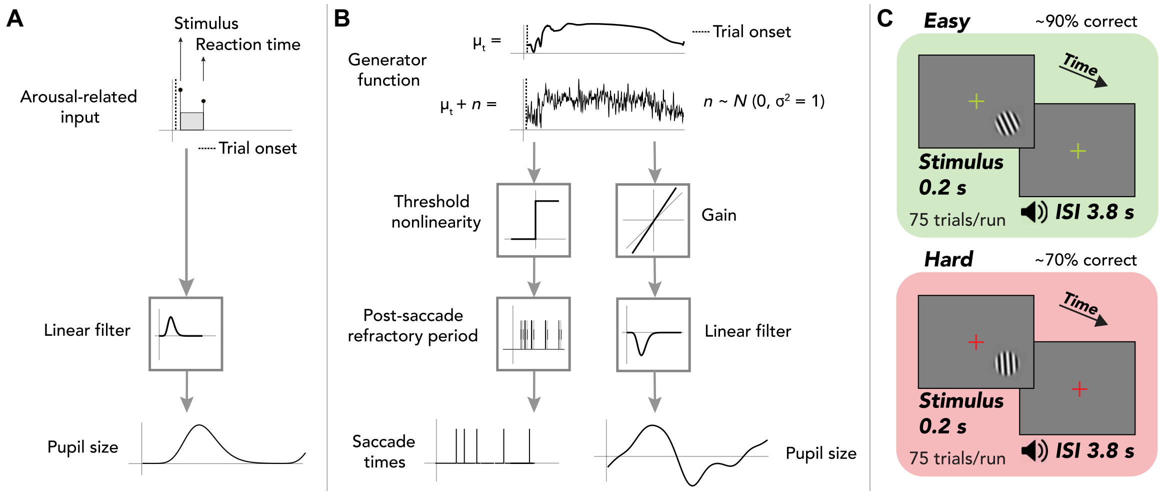

Fig. 1. Models of the task-evoked pupil response and task protocol. (**A**) Consensus model. Arousal-related input (impulses at stimulus onset and reaction time plus a time-on-task boxcar) is filtered by a dilatory low-pass filter, predicting pupil size on a single trial (*17*–*19*, *22*). (**B**) Linear-nonlinear linking model. The generator function, with normally distributed noise (n) added to its expected value (μ_t), is (i) subjected to a threshold and post-saccade refractory period, giving rise to a sequence of saccades, and (ii) multiplied by a gain and filtered by a constrictory low-pass filter, producing pupil size. Generator function, σ reduced by 3× and low-pass–filtered for visualization. Refractory period, gray regions. (**C**) Task, orientation discrimination around vertical. Timing, 0.2-s stimulus presentation, followed by a 3.8-s interstimulus interval (ISI). Observers had to respond during the ISI with a button press and immediately received auditory feedback (correct, incorrect). Design, alternation between separate easy (approximately 90% correct) and hard (approximately 70% correct) runs, 75 trials each. Fixation cross, green or red, indicating easy or hard difficulty. Stimulus, grating (enlarged for visualization).

wait times. The generator function was modulated by the timing of the task and behavioral response, suggesting that the generator function reflects a cognitive representation of task structure. Our model offers a unified explanation for a wide range of phenomena, including (i) entrainment of pupil size and saccade occurrence to task timing, (ii) modulation of pupil response amplitude with task difficulty, (iii) modulation of trial-to-trial pupil noise with task difficulty, (iv) reaction time–dependent modulation of pupil response timing, (v) a constrictory pupil response time-locked to saccades, and (vi) task-dependent distortion of this saccade-locked pupil response.

RESULTS

Experimental design

Human observers (N = 10) performed a forced-choice orientation-discrimination task on a briefly displayed peripheral grating (Fig. 1C). Observers fixated a central cross on a screen while eye position and pupil area were measured. On easy trials, the grating was tilted far from vertical, and observers achieved approximately 93% accuracy on average. On hard trials, the grating's tilt was close to vertical, and observers achieved approximately 70% performance on average. The task difficulty was changed in alternate runs (75 trials per run, 4 s per trial). In a separate version of the task, the wait time between trials was shortened, so each trial was 2 s long. We fit a computational model (Fig. 1B) to measurements of pupil size and saccades (including microsaccades), generating estimates of arousal level.

A correct estimate of arousal should be invariant to trial duration but vary with task difficulty and behavioral performance. Manipulating task difficulty should alter arousal, as one needs to be more alert to achieve good performance on a more difficult task. Small changes in trial duration (i.e., 2 versus 4 s), however, should not substantially alter arousal.

Saccade rate predicted the dynamics of the task-evoked pupil response

Saccade rate, transformed using our linking model, accurately predicted the task-evoked pupil response for different tasks and observers. We included small saccades and microsaccades in all of our analyses (97% were less than 1.5° in amplitude) and will use the term "saccade" from here on to refer to both (see Methods for more details). The probability of saccade occurrence changed over time within a trial (Fig. 2A and fig. S2). It fell after trial onset and rose again thereafter. We assumed that a saccade occurred each time a noisy "generator function" crossed a fixed threshold (Fig. 2B, Methods, and Eqs. 1 and 2). Assuming that the variability in the generator function across trials was standard normal (i.e., indepedent and identically distributed, additive Gaussian noise with SD σ equal to 1), we solved for its expected value across trials (Fig. 2C, Methods, and Eq. 6). That is, the expected generator function was simply a nonlinearly compressed version of saccade rate, where the nonlinearity was an inverse cumulative Gaussian. Next, the expected generator function was mean-subtracted, amplified by a gain, and low-pass–filtered using a linear filter (Fig. 2D, black dashed curve; Methods; and Eqs. 3 and 7), yielding a prediction of the trial-averaged pupil response (Fig. 2E, yellow curve). We estimated gain as the scale factor that resulted in the best match (least squares, i.e., linear regression) to the pupil data. To estimate the linear filter, we measured pupil size following each saccade (Fig. 2D, purple curve) and fit it with a parametric function (Fig. 2D, black curve; Methods; Eq. 8; and fig. S3A). An additive offset, the average pupil size over a run, was added to the prediction of pupil size. For each observer, we modeled the post-saccade refractory period by adjusting the saccade rate function (see Methods). The measured

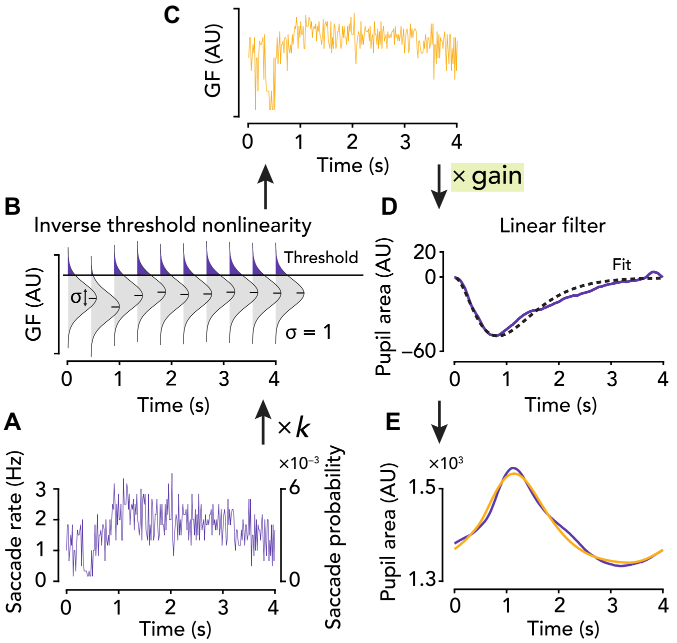

Fig. 2. Model fitting. (**A**) Purple curve, saccade/microsaccade rate in Hz (left axis) or probability (right axis), averaged across trials, low-pass–filtered for visualization. To model post-saccade refractory period, the measured saccade rate function is multiplied by *k*, the shape parameter of a gamma distribution fit to inter-saccade time intervals. (**B**) Inverting the threshold nonlinearity. Gray distributions, standard normal noise (across trials) corrupting generator function ("GF"). Long horizontal black line, threshold. Purple tail, integral equal to the mean saccade rate in (A). Short horizontal black lines, expected value of the generator function, 10 samples only [plotted in full in (C)]. (**C**) Estimate of expected generator function. Low-pass–filtered for visualization. Mean-subtracted and multiplied by gain before filtering. (**D**) Linear filter. Black dashed curve, estimated linear filter—a gamma-Erlang function fit to the saccade-locked pupil response (purple). (**E**) Purple curve, task-evoked pupil response. Yellow curve, model prediction. Gain, highlighted in light green, is the scale factor that gives the closest match between the data and model prediction.

task-evoked pupil response was a close match to the model prediction for all observers (fig. S4; median R^2 across all runs and observers, 81%; mean R^2, 71%; max R^2, 98%; min R^2, 0%; $N = 143$ runs; see the Supplementary Materials for comments on individual differences).

Gain and arousal

Gain, but not offset, scaled with task difficulty

We used our linking model to make predictions of the task-evoked pupil response separately for easy and hard trials. Accuracy in the orientation-discrimination task was much higher for easy than for hard runs, confirming that our manipulation of task difficulty was effective (t test: $t = 39.75$, $P < 1 \times 10^{-6}$, N runs = 143). The data and model predictions were similar in shape and amplitude, and this was true for both easy and hard trials (Fig. 3B and fig. S4). The model's R^2 was not a significant predictor of task difficulty in a mixed effects logistic regression ($t = 1.14$, $P = 0.25$, $N = 143$; see the "Statistical analysis" section in the Supplementary Materials for details).

Arousal should scale with task difficulty because a more difficult task requires greater alertness to achieve good performance. Gain, but not offset, was strongly modulated by task difficulty (Fig. 3, C and D). We fit one free parameter per run, gain. The best-fit gain values, pooled across observers, were significantly larger for hard than for easy runs of trials (t test: $t = 3.38$, $P = 9 \times 10^{-4}$, $N = 143$ runs). This difference remained significant in a mixed effects logistic regression [$F(141) = 10.27$, $P = 1.7 \times 10^{-3}$, N runs = 143] in which we accounted for systematic differences in gain across observers and experimental sessions (see the "Statistical analysis" section in the Supplementary Materials for details). Offset, on the other hand, was not significantly modulated by task difficulty [logistic regression: $F(141) = 0.05$, $P = 0.83$, $N = 143$]. Gain scaled with task difficulty, an expected outcome if gain estimates arousal.

Gain was much higher on error trials

A correct estimate of arousal should vary with behavioral accuracy because arousal influences performance, and because surprise is elicited by tone feedback and confidence influence arousal. We fit our linking model separately to pupil data from correct and incorrect trials (Fig. 3E and fig. S5), expecting that gain would be higher on incorrect trials because of the tone feedback immediately following the button press, which was alarming to observers (particularly on easy runs), and because of the observer's own (imperfect) knowledge of their probability of being correct. Note the timeline of a trial in our task: Observers responded soon after stimulus/trial onset (>50% of responses occurred within 0.4 s). Thus, the pupil's responses to the tone feedback, stimulus, and trial onset were "blurred" together. Because there were far fewer incorrect than correct trials and, hence, fewer saccades and pupil responses to constrain the model estimates, the model fits were understandably poorer overall on incorrect trials (median R^2, 58.16% for incorrect trials; 80.74% for correct; mean R^2: 55.93 and 70.87%). That said, there were many runs with good fits. We used only runs with an R^2 over 50% (81.38% of runs for correct trials and 37.93% of runs for incorrect trials) for this analysis to ensure a fair comparison. Gain was 5.58× higher on incorrect than correct trials (t test: $t = 7.22$, $P < 1 \times 10^{-6}$, N runs = 118 correct, 55 incorrect; Fig. 3E), and this difference was larger in easy than in hard runs (fig. S5, A, D, and E). On the other hand, saccade rate (averaged across observers) was statistically indistinguishable between correct and incorrect trials (cluster-based permutation test: $P > 0.05$ for all time points; fig. S5C). Likewise, the generator function was similar for correct and incorrect trials, particularly before the behavioral report occurred (fig. S5B). "Statistically indistinguishable" comes with a caveat here and throughout: that lack of evidence of a difference cannot be taken as support for our model, but that if we could confirm a difference, it could be taken as evidence against our model, i.e., we attempted to falsify our model and failed to do so.

Gain, but not offset, varied with reaction time

A correct estimate of arousal should vary with reaction time, as arousal is known to influence reaction times. Gain, but not offset, was modulated by reaction time (average across trials in a run), and the direction of modulation was reversed for easy and hard runs (Fig. 3G). That is, the interaction of difficulty and gain was a significant predictor of reaction time [gain × difficulty: $F(139) = 5.80$, $P = 0.017$, $N = 143$; regression analysis; see the Supplementary Materials]. On easy runs, reaction times were faster when gain was higher and slower when it was lower. Vice versa, on hard runs, reaction times were slower when gain was higher and faster when it was lower. The three-way interaction between difficulty, gain, and accuracy was also a significant predictor of reaction time [regression: $F(135) = 6.49$, $P = 0.012$, $N = 143$]. This corresponded with the observation that on the easy runs on which observers had the worst accuracy, the negative relationship between gain and reaction time

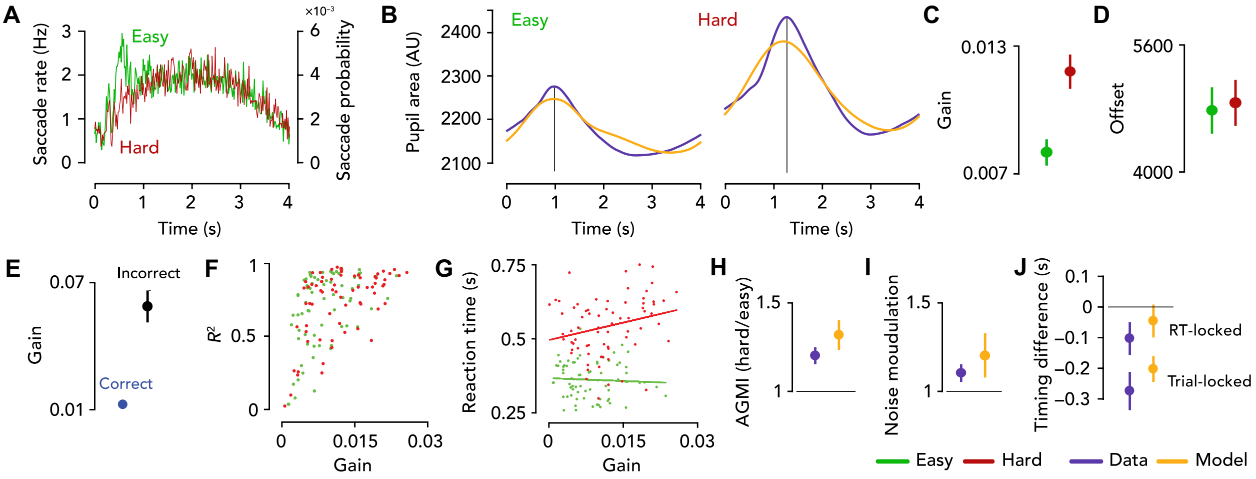

Fig. 3. Model predicts difficulty-dependent modulations in the amplitude, timing, and variability of the task-evoked pupil response. (**A**) Saccade rate was similar for easy and hard trials, except immediately after trial onset. Curves, saccade rate in Hz (left axis) or probability (right axis), average over observers. Green, easy runs. Red, hard runs (the same color convention throughout). (**B**) Task-evoked pupil response was larger in amplitude and more delayed when the task was harder. Purple curve, task-evoked pupil response. Yellow curve, model prediction. Left, easy. Right, hard. Vertical gray line, time-to-peak. 0 s is trial/stimulus onset. (**C**) Gain scaled with task difficulty. Filled circles, mean gain across all runs and observers. Lines, two SEM ($N = 143$). (**D**) Offset did not vary with task difficulty; the same format as (C). (**E**) Gain was much higher for incorrect (N runs = 51) than correct trials ($N = 115$). (**F**) Model's goodness of fit depended on gain (see the Supplementary Materials). R^2 was variable for low gain and concentrated near 1 for high gain. Each circle, one run (N runs = 143). (**G**) Gain was correlated with reaction time and sign of correlation depended on difficulty. Lines, best-fit regression lines. (**H**) Model predicted amplitude modulation of the saccade-locked pupil response with task difficulty. Dots, amplitude or gain modulation index (AGMI), i.e., ratio (hard:easy) of minimum of saccade-locked pupil response or ratio (hard:easy) of best-fit gain. Error bar, two SEM ($N = 15$). Purple, data. Yellow, model prediction. (**I**) Modulation of pupil noise with task difficulty. Dots, ratio (hard:easy) of pupil noise level or of best-fit gain. Error bar, two SEM ($N = 10$). (**J**) Model predicted pupil response timing. Dots, mean difference (easy-hard) in the time-to-peak (s) of the pupil response. Error bar, 2 SEM ($N = 15$). Top, reaction time–locked pupil response. Bottom, trial onset–locked pupil response.

became even more pronounced (than on the runs with higher accuracy). We repeated these regressions but predicting offset instead of gain, and all predictions were nonsignificant [offset × difficulty: $F(139) = 0.06$, $P = 0.80091$; offset × difficulty × accuracy: $F(135) = 0.52$, $P = 0.47$; $N = 143$]. Reaction time depended on gain, as expected if gain estimates arousal level.

An independent prediction: Amplitude modulation of the saccade-locked pupil response with task difficulty matches modulation of gain

Our model predicts that the amplitude of the saccade-locked pupil response scales with gain (fig. S6, B and D). We measured this prediction (Fig. 3H) as the ratio (model:data) of the ratio (hard:easy) of the minimum of the saccade-locked pupil response (fig. S3B) and the ratio (hard:easy) of the mean best-fit gain. A perfect model prediction would yield a ratio (model:data) of 1. The median ratio across observers was 1.13 (mean, 1.11; SEM, 0.09; for observers O1 to O10: 1.06, 0.87, 1.188, 1.15, 1.13, 0.82, 0.78, 1.14, 1.22, and 1.73). We estimated the linear filter used in our model fits from the run-averaged saccade-locked pupil response, throwing away its run-to-run variability. Therefore, this is an independent prediction of the model—that is, a prediction about data that was not used to fit the model parameters.

An independent prediction: Pupil noise scaled with gain

In our model, trial-to-trial variability in pupil size ("pupil noise") is inherited from noise in the generator function. We assume that this input noise is additive and Gaussian, such that when it is amplified by a gain g, its new SD is equal to g (see Methods and Eq. 5). The output of any linear filter acting on a Gaussian input will also be Gaussian (*43*), with the property that if you double the input noise, the output noise will also double (i.e., a linear mapping). Thus, our model predicts that if the best-fit gain is x times larger for hard than for easy trials, then the SD of pupil noise should also be x times larger. Note that this is an independent prediction of the model, as we estimated gain based only on the trial-averaged pupil response, ignoring its trial-to-trial variability.

For each observer, we quantified the level of pupil noise by aligning the pupil responses from each trial and computing the SD of pupil size across trials for every time point from 0 to 4 s (fig. S7A; see the Supplementary Materials, Pupil noise analysis and simulation for details). The pupil noise level was significantly higher for hard than for easy runs for 7 of 10 observers (paired t test, $P < 5 \times 10^{-4}$, N observers = 10). For one observer, O5, the pupil noise level was lower for hard than for easy runs, but this matched their best-fit gain, which was also smaller for hard than for easy. The distribution of pupil noise closely resembled a Gaussian (fig. S7, B and C), validating our noise model and use of SD as a measure of the noise level. The disparity in the pupil noise level between easy and hard runs ("noise modulation") was approximated well by a single scale factor, matching our model's assumption of stationary noise within a trial (fig. S7A). To quantify our model's ability to predict modulation of pupil noise with task difficulty, we computed three ratios for each observer: (1) empirical noise modulation, the ratio (hard:easy) of the average pupil noise level (across time); (2) predicted noise modulation, the ratio (hard:easy) of the best-fit gain (Fig. 3I); and (3) the ratio between empirical and predicted noise modulation. If ratio no. 3 is 1, it means that our model predicted pupil noise modulation perfectly. The median ratio no. 3 (data:model) across observers was 0.95 (mean = 0.89, SEM = 0.09, ratios for each observer O1 to O10: 0.66, 1.09, 0.64, 1.29, 1.06, 1.19, 1.15, 0.84, 0.55, and 0.46). Although noise modulation was slightly larger in the model than data (Fig. 3I and fig. S7A), the predictions were largely in the right direction, and errors were small.

Timing and the generator function

Dynamics of saccade-locked pupil response were modulated by trial duration

The model predicts that the saccade-locked pupil response is different (distorted) from the underlying linear filter because of the threshold nonlinearity for generating saccades. In the data, we found that the nature of this distortion depended on trial duration (i.e., via differences in the generator function) and was particularly bizarre for the 2-s task (fig. S6C), where it caused rippling. Our model closely reproduced this task-dependent distortion (figs. S6, B and D, and S3), perhaps the most idiosyncratic feature of the data, providing a strong piece of evidence in support of the model.

Generator function dynamics were modulated by the timing of the task and behavioral response but largely invariant to task difficulty

We hypothesize that the generator function reflects the timing of the task and behavioral response but is invariant to arousal level. If so, we should find that the generator function changes with trial duration and reaction time but is the same regardless of task difficulty.

We quantified differences in the grand mean generator function on easy versus hard trials (fig. S8) by computing the separation between them at every time point from 0 to 4 s. We chose d′ as a measure of distribution separation, which, in this case (SD = 1; Eq. 1), was simply equal to the difference between the two generator functions. We observed a spike in the generator function on easy but not hard trials at around 0.5 s after trial onset for most observers.

We computed the time of maximal separation as the maximum d′ (over time). O1, O2, O3, O4, O5, O7, and O9 had a large increase in d′ (>1.43) sometime between 0.40 and 0.65 s. For context, the average d′ over the full trial ranged from 0.30 to 0.47 across observers.

To test whether these differences were due to differences in reaction time between easy and hard trials (Fig. 4, A and B), rather than difficulty per se, we conditioned the generator function on reaction time, task difficulty, or both (Fig. 4A). Data were pooled across observers (N = 10). The initial spike in the generator function was larger when reaction time was faster, for both easy and hard trials (Fig. 4A). There was a region of maximal separation (d′) between the generator functions at around 250 ms for both easy and hard trials. We next compared the generator functions (easy versus hard) for fast (Fig. 4C) or slow (Fig. 4D) reaction times separately and found smaller differences. This suggests that "difficulty-dependent" modulations in the generator function (fig. S8) were actually driven by differences in reaction time between easy and hard trials. Therefore, the generator function is modulated primarily by reaction time rather than task difficulty.

The generator function was also modulated by the timing of the task, stretching with trial duration. The generator function sharply rose after trial onset and then slowly fell during the wait time after

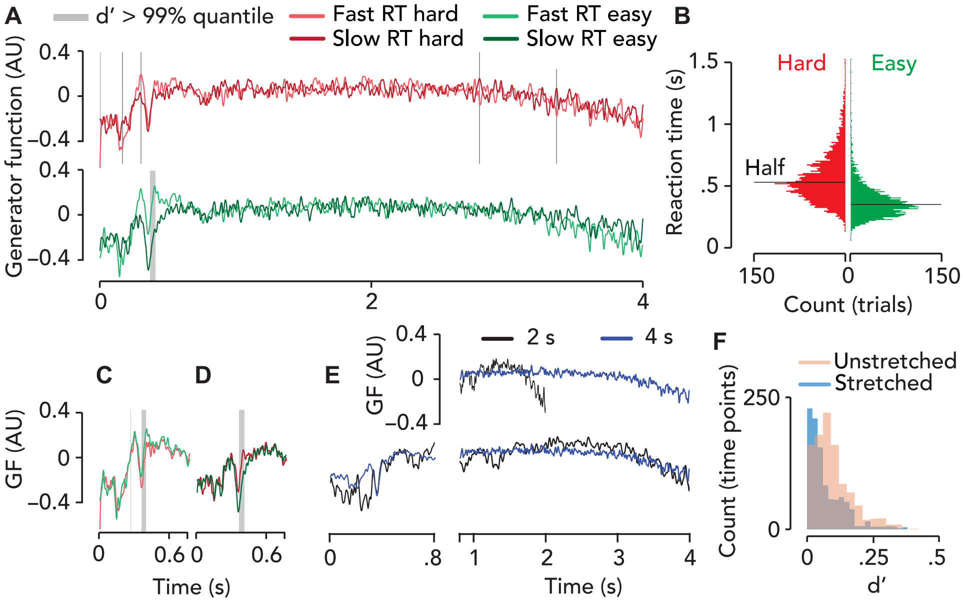

Fig. 4. Generator function was modulated by timing of the task and behavioral response. (**A**) Generator function ("GF") for fast and slow reaction times (median-split). GF computed from average saccade rate (N observers = 10, N trials = 3450) and low-pass–filtered (for visualization). Red, hard runs. Green, easy runs. Light colors, fast RT (below median RT). Dark colors, slow RT (above median RT). Gray regions, times of maximum separation between GFs (i.e., >99% quantile of d′ over time). (**B**) Distribution of reaction times for easy and hard trials (N observers = 10, N trials = 3450). Horizontal black lines, medians. (**C**) GF for fast reaction times (hard versus easy trials), replotted from (A). There were no large separations beyond 0.8 s, so only the first 0.8 s are shown. (**D**) The same as (C), but for slow reaction times. Difficulty-dependent modulations in GF were smaller than reaction time–dependent modulations [compare differences between curves in (A) and (C)]. (**E**) GF for the 2- or 4-s task (N observers = 10; N runs = 97 for 4 s, 39 for 2 s). Black curve, 2 s; blue curve, 4 s. The first 0.8 s are in original time base. Top, unstretched data (i.e., raw data). Bottom, from 0.8 to 4 s, GF from the 2-s task is time-stretched to be 4 s in duration, revealing that the GF is similar for the 2- and 4-s tasks after time stretching. (**F**) Distributions of d′ between the GFs corresponding to the 2- and 4-s tasks, for 0 to 2 s after trial onset (N = 1000 time points). Cyan, d′ when GF for the 2-s task was time-stretched beyond 0.8 s after trial onset. Pink, d′ for raw data (i.e., unstretched).

the button press, i.e., while the observer anticipated the next trial's onset. We hypothesized that the generator function stretches to match the duration of this anticipatory period. Five new observers performed the task with 2- or 4-s-long trials in separate blocks. We compared the generator functions between the 2- and 4-s tasks (Fig. 4E and fig. S11C), with and without time stretching (Fig. 4, E and F). Data were averaged across observers (N observers = 10, N runs = 97 for the 4-s task, 39 for the 2-s task). Specifically, we time-stretched the generator function (Fig. 4E) from the 2-s task to be 4 s in duration. We only time-stretched time points beyond 0.8 s, i.e., after nearly all button presses (approximately 95%) had occurred (Fig. 4B) and when the observer was expecting the next trial to begin. Then, we computed the distribution of d′ across the first 2 s of the trial, between the 2- and 4-s tasks, before and after time stretching (Fig. 4F). d′ was significantly lower for the time-stretched generator function (59% reduction in median; t test: $t = 10.42$, $P = 3.29 \times 10^{-24}$), and this comparison was still significant after log-transforming the skewed distributions to make them more Gaussian ($t = 10.32$, $P = 8.72 \times 10^{-24}$). Note that other reasonable time cutoffs ranging from 0.4 to 1.5 s (corresponding to 57 to 99.6% of button presses having occurred) returned the same statistical outcome ($P < 0.0001$). There were also differences in the generator functions between the 2- and 4-s tasks before 0.8 s (i.e., the never-stretched region, during which stimulus onset, button press, and feedback occur); the maximum d′ of 0.38 occurred 290 ms after trial onset (Fig. 4E).

Pupil response timing was predicted by reaction time–dependent modulations in the generator function

When present, the transient spike in the generator function on easy trials (fig. S8) correctly predicted the advanced time-to-peak of the task-evoked pupil response (Fig. 3B). The task-evoked pupil response peaked earlier for easy than for hard trials (Fig. 3B and fig. S4). We hypothesize that this timing difference is driven by differences in the distribution of reaction times (i.e., advanced timing and lower dispersion for easy trials), which is reflected in the generator function's dynamics (i.e., an early spike in easy but not hard trials). To test this, for each observer, we (i) computed the pupil response time-locked to trial onset ("trial onset–locked," equivalent to task-evoked) or to button press ("RT-locked") (fig. S9, B and D) and computed the difference (easy minus hard) in the response time-to-peak, and (ii) estimated the generator function time-locked to trial onset or reaction time (fig. S9A), used our model to predict pupil size (fig. S9, C and E), and computed the difference (easy minus hard) in the model response's time-to-peak. If our explanation is correct, the model prediction should capture the timing differences (easy versus hard) seen in the data.

For the trial onset–locked pupil response, the timing difference (time-to-peak for easy minus hard) was statistically indistinguishable for model and data (t test: $t = -1.09$, $P = 0.30$, $N = 15$) with an average of −274.3 ms for data (SEM, 62.3) and −202.7 ms for model (SEM, 42.4), i.e., only a 71.6-ms difference between model and data (Fig. 3J). After removing one outlier (O7, who had an abnormal biphasic pupil response, causing the time-to-peak to be very low), the difference dropped to 25 ms. For the RT-locked pupil response, the timing difference was also statistically indistinguishable between model and data (t test: $t = -0.81$, $P = 0.43$, $N = 15$), with an average of −102.8 ms for data (SEM, 53.6) and −46.7 ms for model (SEM, 55.4), i.e., a 56.1-ms difference between model and data. After removing the outlier (O7), this difference dropped to 10.1 ms. Thus, timing differences in the task-evoked pupil response were driven by a spike in the generator function present on easy but not hard trials, which was linked to faster reaction times. When the spike was also present on hard trials (i.e., in the RT-locked generator function), these timing differences disappeared. This demonstrates that pupil response timing reflects reaction time–dependent modulations in the generator function.

Model comparison

Pupil response amplitude depends jointly on gain and generator function

The amplitude of the task-evoked pupil response, commonly used as a heuristic measure of arousal level, depends on both gain and the dynamics of the generator function. We separately quantified the components of pupil response amplitude (i.e., maximum pupil size relative to trial onset baseline) due to the gain and generator function by analyzing our model's prediction before and after gain modulation for easy versus hard runs (fig. S10). The pregain model prediction captured some of the differences in pupil response amplitude between easy and hard runs both for individual observers (fig. S10A) and, on average, over all observers (fig. S10B). However, the difficulty-dependent change in the generator function and the difficulty-dependent change in gain were both needed to account for the pupil size measurements.

Comparison with existing models

Our model generalized between tasks with different trial durations, whereas the consensus model could not. To assess generalization performance, we had a group of observers ($N = 5$) perform two versions of the task, which differed only in the duration of the waiting time between trials (2 or 4 s). Pupil and saccade data were markedly different for the two trial durations (fig. S11, A and B), but behavioral accuracy was nearly identical (average accuracy across runs, 2- versus 4-s task, hard: 72.93% versus 72.46%; easy: 96.67% versus 98.78%; $t = 0.82$, $P = 0.42$). A correct model linking arousal and pupil size should recover similar parameter estimates of arousal for both tasks despite the different pupil size dynamics. Therefore, we tested whether the parameters estimated from one dataset applied to the other nearly as well, generating accurate predictions of pupil size. To do so, we fit either our model or the consensus model to the 2-s data and evaluated the best-fit parameters on the 4-s data, and vice versa. This was done separately for easy and hard runs for each observer. We quantified generalization performance as the median reduction in R^2 between the out-of-sample prediction versus in-sample fit. For our model, we used the average gain across all easy or hard runs for the out-of-sample prediction. For the consensus model, we fit the amplitudes of "arousal-related" impulses at trial onset and button press and of a "time-on-task" boxcar (fig. S11D). We used these three amplitude parameters (averaged over easy or hard runs) for the out-of-sample prediction. For our model, the median reduction in R^2 across observers was 14.67% (reductions for each observer, 4 → 2-s prediction: 10.06, 17.21, 0.60, 36.42, and 19.94%; 2 → 4 s: 12.14, 20.96, 5.40, 7.01, and 124.66%). For the consensus model, the median reduction in R^2 across observers was 46.89% (reductions for each observer, 4 → 2-s prediction: 47.07, 65.65, 62.33, 16.94, and 59.49%; 2 → 4 s, 21.645, −16.91, 117.31, 7.49, and 46.73%). For our model, the median R^2 values for the in-sample fits were 86.27% (2 s) and 78.62% (4 s). For the consensus model, they were 94.16% (2 s) and 39.14% (4 s). Together, these results suggest that the consensus model overfits and that its predictions are not self-consistent.

The primary reason that the consensus model fit the 2-s data better than the 4-s data was the presence of a large pupil constriction relative to pretrial baseline around 1 s after trial onset in the 4-s but not the 2-s task (fig. S11A) (*21*). The consensus filter is dilatory and so cannot capture such constrictions at all (fig. S11D). On the other hand, our model predicted this large constriction on the 4-s data and the much smaller constriction in the 2-s data (fig. S11A). The reason is simply the prolonged generator function (fig. S11C) and, hence, prolonged input to the constrictory linear filter. It is clear a priori that adding a second linear system with a constrictory filter evoked by the button press, as some have proposed (*21*, *22*), would make the consensus model's in-sample fit in the 4-s task much better, but it would also make the model even less generalizable. This is because, to explain the data, the impulse amplitude of the constrictory response would have to be near zero for the 2-s task but larger and negative for the 4-s task.

Pupil size is a linear transformation of the generator function, not trial events

Model comparisons were performed to assess whether the pupil is a shift-invariant linear transformation of impulses time-locked to trial events, as existing models assume, or of the gain-modulated generator function, as our model assumes. To test these assumptions, we computed the ratio of the best-fit scale parameters (i.e., gain or impulse/boxcar amplitudes) from the 2- and 4-s tasks. We refer to this ratio as the additivity/shift invariance or "ASI" index. If pupil size is a linear shift–invariant transformation of its input, this ratio should be 1. For the consensus model, the mean ASI index across observers was 2.45×10^{11} (SEM, 1.60×10^{11}) for the boxcar, 1.54×10^{10} (SEM, 1.54×10^{10}) for the trial onset impulse, and 0.61 (SEM, 0.13) for the button press impulse. The medians were 0.53, 0.57, and 0.58, respectively. The reason for the large mean values is that the amplitude estimates were unstable because the three parameters traded off with each other. For our model, the mean ASI index across observers was 0.9323 (median, 0.80; SEM, 0.12), i.e., closer to 1. Next, we computed a homogeneity/shift invariance or "HSI" index. This was the ratio (4 versus 2 s) of the quotient of the hard and easy best-fit gain values (our model) or amplitudes (consensus model). If pupil size is a linear shift–invariant transformation for its input, the HSI index should be 1. For the consensus model, the mean HSI was 0.09 (SEM, 0.06) for the boxcar, 1.30×10^{11} (SEM, 1.30×10^{11}) for the trial onset impulse, and 6.06×10^{10} (SEM, 6.06×10^{10}) for the feedback impulse. The medians were 0.01, 3.56, and 2.16, respectively. For our model, the mean HSI index was 1.13 (median, 1.17; SEM, 0.18), i.e., closer to 1. Together, these results suggest that the pupil is a shift-invariant linear transformation of the gain-modulated generator function, not trial events. These results also show that gain was robust to small differences in the wait time between trials, as one would expect of an estimate of arousal.

DISCUSSION

We propose a model of the task-evoked pupil response, which capitalizes on the hypothesis that a common input drives both saccades and pupil responses, i.e., the "common drive hypothesis" (*12*, *16*, *25*–*27*). Our model estimates two main components, the generator function (i.e., the common input) and gain. We hypothesize that the generator function reflects task timing and gain reflects arousal. To test this idea, we fit our model with saccade and pupil size measurements from a visual orientation-discrimination task in which task difficulty and trial duration were varied. The model accurately predicted pupil size from saccade rate. Estimates of gain were modulated by task difficulty and behavioral accuracy, but not by trial duration. On the other hand, the estimated generator function was modulated by the trial duration, but not by difficulty or accuracy. Both the gain and generator function were modulated by reaction time.

Existing models assume that pupil size is a linear transformation of arousal-related impulses at the time of trial events. Such models failed to generalize across trials with different timing despite achieving nearly perfect in-sample fits, demonstrating that this assumption is false and that these models are overfitting. On the other hand, our model achieved good generalization and fits. One implication is that it is necessary to measure both saccades and pupil size to correctly estimate the input to the pupil and to thereby constrain estimation of arousal.

Our model provides a unified explanation for a number of seemingly disparate observations, including entrainment of pupil size and saccade rate to task timing, amplitude modulation of pupil responses with task difficulty, modulation of trial-to-trial pupil noise with task difficulty, a constrictory pupil response time-locked to saccades, task-dependent distortion of this saccade-locked pupil response, and reaction time–dependent modulations of pupil response timing. These results provide strong support for the hypothesis that a common input (the generator function) drives saccades and pupil size, a saccade occurs when the generator function crosses a threshold, the task-evoked pupil response is a gain-modulated and low-pass–filtered transformation of the generator function, and arousal corresponds to the gain.

Over the past 80 years, measuring pupil size has become a popular tool for inferring arousal level (*16*). Inexpensive, noninvasive, and high-fidelity pupil size recordings have many advantages over the alternatives, i.e., measurement of spiking activity in neuromodulatory loci (*8*, *12*), whole brain imaging (*15*), and in vivo biochemical recordings (*44*). A large number of recent publications have linked pupil size with behavior (*17*, *21*, *45*–*47*) or neural signals (*11*, *12*, *14*–*16*, *26*). Approaches to estimating arousal from pupil size fall into two categories: heuristic and model-based. A popular heuristic analysis in systems neuroscience involves computing the correlation (or regression) between the time course of pupil size and neural activity in some population of interest, where pupil size is treated as a measurement of arousal level (*14*, *15*). The implicit assumption is that the neural input to the pupil tracks with arousal level and that delays and temporal summation caused by the sluggish response of the iris muscles (*48*, *49*) do not substantially obscure the underlying correlation between the neural signals. Another popular heuristic approach estimates arousal level based on the height of the pupil response (i.e., max or max-min size) (*50*).

The most popular model-based approach is a linear systems model, which assumes that the pupil response is a low-pass–filtered transformation of arousal-related neural activity aligned to task events (*17*–*20*, *22*, *29*, *51*). This assumption is often justified on the basis of the physiological findings that LC neurons spike at the time of salient events (*8*) and that LC activity or microstimulation precedes pupil dilation (*11*, *12*). Some highly cited papers thus claim that it is possible to infer activity in the LC-NE system from pupil size (*52*). This consensus model is nearly identical to one that has been used in functional magnetic resonance imaging (fMRI), in which the strength of neural activity is inferred by regressing a prediction, the convolution of the hemodynamic response function with impulses at trial event times, on the fMRI signal. In fMRI, this approach was validated by the observation that the fMRI signal approximates a shift-invariant linear transformation of local spiking activity (*53*) and that neural activity is time-locked to stimulus onsets in many sensory areas of the brain. However, linearity of pupil size with respect to its purported input has never before been tested. Furthermore, it is unclear whether the neural input to the pupil is aligned with trial events. Our model comparisons suggest that both assumptions are false—the consensus model's parameters did not generalize across trials with different durations, demonstrating that pupil size is not a linear transformation of an input time-locked to trial events. Earlier versions of the consensus model (*23*, *24*) assumed that the input to the pupil were impulses that could occur at any time (analogous to our generator function) rather than being time-locked to task events. We found, however, that the parameters of such models were not recoverable (see the "Parameter recovery: Deconvolution method of Wierda and colleagues" section in the Supplementary Materials), suggesting that these models are underconstrained for pupil size measurements alone.

The challenge of estimating arousal from the task-evoked pupil response is that the input to the pupil and arousal level are both unknown, making the problem underconstrained. The consensus model constrains the problem by assuming that the input's timing is equal to the task timing (which is known because it is set by the experimenter) and that the input's form is equal to some particular parametric form (e.g., impulse or boxcar). We adopted a different approach in which we used a model to estimate the input to the pupil from measurements of saccades rather than assuming its timing and form. By analogy, fMRI measurements in early visual cortex can be analyzed by adopting a model that maps from the pixel intensities of an image to the evoked neural activity (*54*); such models are based on the known anatomy and physiology of the brain and include multiple stages comprising nonlinear and linear operations (*55*). The advantage of such "image-computable" models is the ability to make predictions of a measurable signal (i.e., the fMRI signal) from another measurable physical signal (i.e., an image) and to compare the prediction to data under many conditions, putting a constraint on the possible operations in between. Here, we took a step toward adopting an analogous theoretical framework for estimating arousal from pupil size. Future work is needed to develop a model that links task structure to the generator function. Currently, this mapping is unknown. If this can be achieved, then it will be possible to combine our model with this one, yielding a model that predicts pupil size from the timing of task events, i.e., a "task-computable model."

We speculate that the generator function, the driver of saccades and pupil size, is influenced by multiple aspects of cognition. We manipulated trial duration in our study and found that during the interstimulus interval (ISI), the generator function stretched with trial duration (Fig. 4E), whereas between the stimulus onset and behavioral response, it was modulated in a more complex way (Fig. 4E and fig. S11C). It is well known that trial duration influences task engagement and anticipation (*56*, *57*), as well as uncertainty (if duration varies). Thus, modulations in the generator function may reveal the extent of task engagement and anticipation. In our model, the generator function is simply a transformed version of the trial-averaged (micro)saccade rate. Thus, previous results revealing that (micro)saccade rate is modulated by temporal expectation and attention (*38*, *41*) also support this conclusion. Extending this logic, it is well accepted that many cognitive processes can affect (micro) saccade rate, including attention, working memory, and voluntary saccade planning (*58*, *59*). Thus, although we only tested the effect of trial duration on the generator function in this study, we expect that the generator function reflects many other aspects of cognition.

Our results demonstrate that the amplitude of the task-evoked pupil response depends on both the generator function and gain (fig. S10), implying that the same measured pupil response can be due to many different combinations of the two components. Hence, we used model fitting and comparisons to tease out whether pupil size modulation was primarily due to changes in gain, changes in the generator function (i.e., the input to the pupil), or both. These results prompt a reinterpretation of many previous findings relating the amplitude of the task-evoked pupil response to various psychological processes, including memory, learning, cognition, attention, and decision-making. Our model makes a strong prediction that can aid in this reinterpretation: Psychological processes that influence both saccade rate and pupil size do so via the generator function. Processes that influence pupil size, but not saccade rate, do so via gain (*37*). Some processes will influence both. Our model is needed to quantify these contributions. Another prediction of our model is that any variable that affects saccade rate will also affect pupil size. The inverse statement that any variable affecting pupil size will also affect saccade rate is not implied by our model, and experimental evidence speaks against it. For example, arousal influences pupil size but not microsaccade rate (*59*).

Physiological basis

We can only speculate about the physiological basis of the four main components of our model: the generator function, threshold, gain, and linear filter. We offer a disclaimer that our model is simple and abstract by design. Multiple model components could correspond to one part of the neural circuitry, or vice versa; one model component could correspond to multiple parts of the circuitry. Further anatomical, physiological, and psychophysical experiments are needed to determine these correspondences.

We speculate that the generator function corresponds with the pooled activity of SC neurons. Specifically, we hypothesize that (i) SC drives both saccades and pupil responses, and (ii) SC encodes a cognitive representation of the task, including its timing. Support for the first of these hypotheses is evidence that the intermediate/deep layers of the SC (SCi) are a common (indirect) input to the muscles that control pupil responses and saccades (*12*, *16*, *25*–*27*). SC projects directly to the preganglionic EW nucleus, which controls pupil constriction and is most widely known for its role in mediating both the pupil's accommodative near response and light reflex (*26*, *43*, *48*, *60*). SC, most known for its role in saccade generation (*61*), is involved more generally in orienting behaviors (including eye, head, and body movements) that guide target selection and spatial decisions (*62*). However, spiking in SC neurons is also locked to a pupil response (*12*, *25*, *26*), and crucially, microstimulation of SC neurons (above or below the threshold for saccade generation) evokes a pupil response (*12*, *25*, *27*).

Support for the second of the above two hypotheses is evidence that SC neurons encode the saliency of visual, auditory, and somatosensory events (*63*), thus allowing it to build a timeline of a task (i.e., reflecting task timing). SCi neurons respond to behaviorally relevant stimuli in a manner that is substantially invariant to their particular retinotopic location (*64*). Optogenetically inhibiting SC, even ipsilaterally to a visual stimulus, causes increases in psychophysical thresholds in a detection task and increases in lapse and guess rates, suggesting that SC activity is important for maintaining task engagement (*65*). Muscimol-mediated inactivation of SCi neurons causes biases in target selection, for both saccadic and button press responses, despite not causing generic impairment of movement, demonstrating that SCi encodes target priority in an effector-independent manner (*62*). We speculate that a cognitive representation of task structure generated in higher cognitive areas, including dorsolateral prefrontal cortex and anterior cingulate cortex, modulates SC activity, which, in turn, controls pupil size. This is consistent with previous proposals (*16*, *26*, *31*).

We speculate that the threshold in our model corresponds to the aggregate behavior of the circuit connecting SC, omnipause neurons, the brainstem burst generator, and the extraocular muscles. This circuit is known to initiate saccades and maintain eye position in the orbit (*58*). The threshold in our model may approximate this entire circuit's input-output relation—i.e., an abstraction of SC-driven saccade initiation. Thresholding of neural activity in SC or frontal eye field is likely the most influential model of saccade initiation (*66*, *67*).

We speculate that the gain in our model corresponds with neuromodulation of activity in the preganglionic EW. A possible circuit-mediating gain modulation is the direct inhibitory noradrenergic projection from LC onto EW (*2*, *26*, *68*) found in cats. However, the existence of this connection in humans is controversial (*69*). There are also direct projections from LC onto SCi, which then, in turn, projects to EW both directly and indirectly (via the mesencephalic cuneiform nucleus) (*16*, *26*). The picture is enriched by the fact that cat EW in fact receives projections from many other neuromodulatory nuclei, including the ventral tegmental area, the pars reticulata of the substantia nigra, and the raphé nuclei (*68*, *70*). If these same afferent connections exist in humans, gain may be controlled by a number of neuromodulators, not just by the LC-NE system. This would explain why pupil responses have been observed locked to microstimulation of neurons in noradrenergic, serotonergic, and cholinergic brainstem loci (*11*–*13*, *31*). One could test this idea by measuring EW activity and pupil size with pharmacological manipulations that selectively affect these neuromodulators. If drugs that selectively affect the dopamine system, for example, gain-modulate EW activity, this would provide evidence that the pupil's gain can be influenced by a number of neuromodulators, which collectively define "arousal level" (*31*). This would also explain why pupil responses are modulated by reward and reinforcement learning (*21*), which are commonly associated with the dopaminergic system.

We speculate that the linear filter in our model corresponds with the activity of the iris constrictor muscle driven by the parasympathetic nervous system. In the parasympathetic pathway, the constrictor muscle is driven by cholinergic neurons arising in the ciliary ganglion, which are driven solely by afferents from EW (*71*). In the sympathetic pathway, the iris dilator muscle is under the control of adrenergic neurons in the superior cervical ganglion, which is driven by the intermediolateral nucleus of the spinal cord (*16*, *72*, *73*).

Pharmacological and physiological studies reveal that pupil responses during task performance and rest are primarily under parasympathetic control and that arousal modulates this circuitry. The task-evoked pupil response is nearly eliminated by a local parasympathetic antagonist (i.e., applied to the eye) but unaffected by a local sympathetic agonist (*74*), demonstrating that the task-evoked pupil response is primarily under parasympathetic control. Likewise, a local sympathetic antagonist or agonist has no effect on pupil size fluctuations at rest, while a local parasympathetic antagonist attenuates these fluctuations (*4*, *5*). Enriching this account, an orally administered central alpha-2 agonist (i.e., decreases central NE) decreases pupil size fluctuations (at rest) in the light but has no effect on pupil size in the dark. By contrast, an alpha-2 antagonist (i.e., increases central NE) increases pupil size fluctuations in the light and has no effect in the dark (*33*, *34*). Together, this suggests that global neuromodulation (e.g., related to function of LC) can interact with the pupillary light reflex, which is primarily driven by the parasympathetic pathway (*75*, *76*). Supporting this idea, very high or low arousal attenuates the constrictory pupil response to a transient luminance change (*1*, *5*, *35*, *74*, *75*, *77*). In an awake cat, a startling auditory tone leads to a characteristic rising-and-falling pupil response even if the sympathetic pathway is surgically removed, although its amplitude is attenuated by 1.5×, and its dynamics are more low-pass compared to a normal pupil (*32*). Stimulation of cortex or hypothalamus of a sympathectomized animal causes inhibition of the pupil light reflex, and stimulation of superior cervical ganglion in an intact animal does not (*32*, *75*). Likewise, in macaque, the modulatory influence of arousal level on the pupillary light reflex is nearly the same in a sympathectomized and intact eye (*75*). These results support the hypothesis that inhibition of the parasympathetic pathway is the primary driver of the dilatory portion of the task-evoked pupil response (*32*). One study of the light reflex found that simply passing the spike train of an EW neuron through a constrictory low-pass filter leads to an accurate prediction of pupil size (*48*), suggesting that the relation between EW activity and pupil size is linear. This supports our speculation about parasympathetic control of the linear filter in our model. Regardless of what drives activity in EW, whether it is light, accommodation, sound, self-generated movement, or increased cognitive engagement, the pooled magnitude of EW activity (via ciliary ganglion) determines the total number of acetylcholine molecules released onto the iris constrictor muscles and, hence, the magnitude of pupil constriction. Additional physiology experiments that test linearity directly by simultaneously measuring pupil size and its neural inputs (i.e., the inputs to the dilator and constrictor muscles) are needed.

Our model assumes that pupil response variability is driven by the noise in the generator function and scales with gain. While previous studies have differentiated between variability in pupil responses at rest and during task performance (*1*, *4*, *32*), our model suggests that these share the same origin and are simply points along a continuum defined by the dynamics of the generator function. At one extreme (spontaneous), the generator function is flat in expectation, so saccades and pupil size are unpredictable (fig. S1, A and B). This input may be systematic but endogenous (e.g., because of interoception or mind wandering) and therefore unpredictable to an outside observer and cancels out when averaging over time. At the other extreme (task-evoked), the generator function entrains to the timing of external events (e.g., trial onsets), so saccades and pupil size modulations are more predictable (fig. S1, C, D, and F). Pupil fluctuations are of a similar amplitude during task or rest (fig. S1E). The assumption here is that the same (primarily parasympathetic) circuitry controls both task-evoked and spontaneous pupil fluctuations (*5*, *74*). Previous studies support this idea. Pupil noise in the two eyes is almost 100% correlated, implicating a noise source that drives both eyes, i.e., EW or higher in the circuitry (*43*). Confirming this, pupil noise is attenuated if the nerve between the ciliary ganglion and its input, EW, is cut (*1*). Fentanyl, an opioid agonist that attenuates inhibition of midbrain structures (like EW), depresses equiluminant pupil size fluctuations, suggesting that inhibition of EW produces these fluctuations (*36*). This is consistent with a physiological study showing that EW has a high tonic firing rate, which is inhibited by a transient step-up in light level (*48*)—i.e., arousal may also inhibit tonic activity in EW. A statistical analysis by Stanten and Stark (*43*) revealed that pupil noise is multiplicative and Gaussian. This can arise from an additive Gaussian noise source followed by a gain and then subjected to a linear filter (*43*, *78*–*80*). Stanten and Stark found that the SD of pupil noise scales with both light level and fixation depth (accommodation), suggesting that the same noise source is shared among all parasympathetic-mediated pupil responses. Consistent with this, we found that arousal modulated both gain and the SD of pupil noise (fig. S7), bolstering the idea that the task-evoked pupil response is parasympathetic-mediated, and its noise is inherited from the gain-modulated generator function. Stark concluded that this noise source must be in the EW or higher (*43*, *78*–*80*). Our results demonstrate that variability in saccade generation and pupil responses is shared, i.e., we hypothesize that pupil noise must originate above EW, possibly in SC.

Possible confounds

The saccade-locked pupil response is neither a luminance or accommodation response nor an artifact of foreshortening caused by a change in eye position (a potential source of error in infared eye trackers). Foreshortening error occurs when the detected pupil area or diameter is smaller (i.e., a squashed ellipse versus a circle) because of gaze direction deviating from screen center (i.e., where the eye tracker is angled). The saccade-locked pupil response occurs with a delay of 0.9 s and takes approximately 4 s to come back to baseline. An artifact based on eye position should arise and dissipate much more rapidly than 4 s. Moreover, many of the saccades measured in our study are corrective (i.e., they are toward the fixation point, correcting for drift). These small corrective saccades correspond to a relatively large decrease in pupil size, as opposed to the small increase predicted by a reduction in foreshortening. On the other hand, real pupil size changes can cause erroneous estimates of gaze position (*81*, *82*). However, a multisecond change in pupil size would not cause erroneous detection of a saccade, suggesting that this artifact cannot alter our conclusions (*82*). It has been suggested that the saccade-locked pupil response occurs because of a transient change in luminance (*29*, *83*). Stark ruled out this explanation in his 1966 paper (*28*) by varying luminance (i.e., showing a flash of darkness) at variable delays with respect to the time of a voluntary eye movement and observed that the saccade-locked pupil response effectively quashed the luminance response when they were evoked simultaneously. Furthermore, if the saccade-locked responses were luminance-evoked, the shape and amplitude of saccade-locked responses should be the same during hard and easy runs of our task because the overall luminance was the same in each. However, the saccade-locked pupil response had a significantly higher amplitude during hard than during easy runs. This could not be explained by differences in saccade amplitude, which might affect pupil size by modulating retinal illumination (see the Supplementary Materials). In addition, the shape of the saccade-locked pupil response changed markedly between the 2- and 4-s tasks (figs. S3 and S6C). This task timing–dependent deformation was predicted quantitatively by our model (fig. S6, A, B, and D) but is not predicted by existing models of the light reflex. Others have suggested that the saccade-locked pupil response is caused by some visual change other than luminance (*84*, *85*), but again, this does not explain why the amplitude of the saccade-locked pupil response scaled with task difficulty despite the stimulus being the same. Stark suggested that the saccade-locked pupil response was due to changes in accommodation arising from switching fixation locations. The eye movements in his study were 20° in magnitude. This explanation does not make sense for the small fixational saccades and microsaccades that were analyzed in our study (see the "Saccade detection and rate estimation" section), when depth was virtually identical for different fixation locations. Furthermore, microfluctuations in accommodation during fixation do not predict fluctuations in pupil size (*60*), and accommodation changes during task performance are too small to explain the task-evoked pupil response (*86*). The function of the saccade-locked pupil response, if there is one, remains unclear (*47*); however, its origin is clearer. It does not seem to be visually or accommodatively evoked [although see (*83*)]. Its causal link to activity in SCi (*12*, *25*, *27*) is the strongest evidence we currently have about its control.

Capabilities and limitations

Our method is applicable to eye data collected in a broad array of tasks. While we used fixed trial durations and a blocked design (different difficulty levels), our algorithm is readily extendable to tasks with event-related designs (with jittered ISI durations across trials), tasks with interleaved trial types (see our analysis of error trials), and tasks with multiple overlapping sets of conditions. One limitation of our method is that it should be used with equiluminant stimuli to avoid eliciting pupil light reflex responses that might drown out the smaller arousal- and cognitive-related responses or otherwise confound estimation of arousal. To overcome this, it may be possible to model the pupil light reflex (a linear system) based on the luminance of the display and first remove its influence from the pupil time series and then apply our model. Another limitation is that our model assumes that gain does not change within a trial. We suspect that this is approximately true—that the gain changes slowly—because pupil size changes slowly. Furthermore, our model predicts pupil dynamics well, assuming a static gain. Future adaptations of our method can perhaps introduce additional assumptions to estimate gain on a finer time scale. One last limitation: Our model predicts pupil size from saccade rate. It is impossible to reverse this and estimate saccade rate from pupil size because the linear low-pass filter destroys fine temporal structure in the generator function.

In conclusion, it has long been known that the size of the pupil reflects cognition and arousal (*1*, *7*, *16*, *21*, *47*). However, it was found only recently that pupil size and saccades share a common neural input (*12*, *25*, *27*), prompting a revision of existing models of the task-evoked pupil response. In the model we propose here, a common, noisy input (i.e., generator function) drives both pupil size and saccades, and a gain subsequently modulates pupil size. We find that the common input reflects a cognitive representation of task structure (including its timing) and that the gain reflects arousal level. We offer an algorithm, based on our model, that estimates both components from measurements of pupil size and saccades. Our code toolbox is available at https://github.com/csb0/PCDM and our experimental data and code are available at https://archive.nyu.edu/handle/2451/63809.

METHODS

Observers

Observers (N = 10, 6 females, 4 males) were healthy human adults with no known major neurological disorders and with normal or corrected-to-normal vision. Five observers (O1 to O5) participated in the 4-s task. Five additional observers (O6 to O10) participated in both the 2- and 4-s tasks. All observers were naive to the purposes of the experiment except O1 and O6 (two of the coauthors). O2, O3, O4, O7, and O9 had little to no experience participating in visual orientation-discrimination experiments, while the remaining observers had considerable experience. Experiments were conducted with the written consent of each observer. The experimental protocol was approved by the University Committee on Activities Involving Human Subjects at New York University.

Equipment and setting

Stimuli were displayed on a cathode ray tube monitor (22″ diagonal; Hewlett-Packard p1230) with a refresh rate of 75 Hz and a resolution of 1152 by 870, custom-calibrated for gamma correction. The task and visual stimuli were controlled with MATLAB software based on the MGL toolbox (gru.stanford.edu/mgl). The observer's eyes were 57 cm from the monitor, and head was fixed in a chin-and-forehead rest, which minimized head movement. Pupil area was recorded continuously during task performance using an Eyelink 1000 infrared eye tracker (SR Research Ltd., Ontario, Canada) with a sampling rate of 500 Hz. Nine-point calibration and validation was performed before each run of 75 trials to ensure proper measurement of eye position. The room was dark, and the door remained shut for the length of the experiment, ensuring that dark adaptation was not interrupted. Cell phones were taken away to prevent vibrations and light emissions that might alter engagement or arousal.

Task design

Observers performed a two-alternative forced-choice visual orientation-discrimination task, which varied in difficulty level in alternate runs of trials. We instructed observers to report whether a small, peripheral grating was tilted counterclockwise or clockwise of vertical by pressing the "1" or "2" keys on a keyboard with the nondominant hand. Auditory feedback provided immediately after button press indicated accuracy (high tone, correct; low tone, incorrect). Observers were asked to fixate a central cross (width, 0.7°) and to minimize blinking (average blink rate, 0.13 Hz; range across observers, 0.04 to 0.54 Hz). The (equiluminant) color of the fixation cross was red on hard runs and green on easy runs to explicitly inform the observer of the current task difficulty. We instructed observers to be as accurate and as fast as possible in responding and to maintain fixation at all times.

The experiment consisted of two sessions conducted on two separate days. Each session comprised five separate runs (75 trials per run), alternating between easy and hard. Therefore, observers performed either three easy runs for the first session and three hard runs for the second session, or vice versa. This order was randomly chosen. Five of ten observers did two easy runs in their first session. Observers could rest between runs for as long as they needed but typically rested for less than 1 min. Each run consisted of 75 4-s trials, lasting for approximately 5 min each, for a total of approximately 60 min per session (with breaks).

Each trial consisted of a stimulus presentation of 0.2 s followed by an ISI of 3.8 s in the 4-s version of the task or 1.8-s in the 2-s version. The observer could only respond during the first 1.8 s of the ISI. If observers missed this response window, no tone would play, indicating a missed trial. Observers missed the response window exceedingly rarely (15 of 6900 trials, i.e., 0.22%).

Before beginning the experiment, observers were trained initially for 75 or 125 trials, depending on their familiarity with orientation-discrimination tasks. Orientation threshold was measured as the staircase value on the final training trial and was used to set the initial stimulus tilt in the first hard run of trials in the experiment.

Stimulus

The stimulus was a grating (diameter 1.5°; spatial frequency, 4 cycles per degree; contrast, 100%) multiplied by a circular envelope (diameter, 1°) with raised-cosine edges (width, 0.25°). The stimulus had a mean luminance (over space) that was equal to the background luminance (mid-gray), so that it would not be expected to evoke a luminance response from the pupil, driven by intrinsically photosensitive retinal ganglion cells with large spatial integration areas [i.e., wide dendritic trees (*87*)]. The stimulus was presented in the lower right hemifield, 5° from the center of the screen (Fig. 1C). On easy trials, the tilt of grating was fixed at ±20° from vertical, yielding 93% correct discrimination accuracy on average across observers. On hard trials, the tilt of the grating was controlled adaptively according to prior performance. Specifically, the absolute value of the tilt was controlled by two interleaved two-down–one-up staircases with initial thresholds of 5° and 0°, respectively, and an initial step size of 1, which converged to 70% discrimination accuracy (on average). The tilt's sign (i.e., clockwise or counterclockwise of vertical) on each trial was determined randomly. The staircase value (stimulus tilt) on the final trial of a run was carried over to the following run or session.

Pupil data preprocessing

Pupil size was recorded as the area of a model ellipse in the arbitrary units (AUs) specified by the Eyelink eye tracker's firmware. The Eyelink units are proportional to pupil area (mm^2) (*88*). Therefore, it can be easily used in regression models, and if the pupil size (AU) in one condition is doubled compared to another condition, that relation holds in physical units. Gaze position and pupil area time series were linearly interpolated during and 150 ms before and after blinks, following (*45*).

To estimate the saccade-locked pupil response, we first band-pass–filtered the pupil area time series to remove low frequencies that would confound deconvolution and high frequencies that were not physiologically plausible (i.e., measurement noise). For all other analyses (e.g., for data analysis and model fitting), we just low-pass–filtered the pupil size signal to remove high-frequency measurement noise. The band-pass filter was a Butterworth fourth-order zero-phase filter ("filtfilt" in MATLAB) with 0.03- and 10-Hz cutoffs. The low-pass filter was a Butterworth second-order zero-phase filter with a 10-Hz cutoff.

We used a custom convolution boundary handling method to ensure that we did not generate large, artifactual signals at the edges of the filtered time series. Boundary handling was performed differently for the beginning and end of the time series. The first sample of the time series was repeated N times and concatenated to the front of the time series. The last N samples were mirrored and concatenated to the end of the time series. N refers to the length of signal. This method was used because the dynamics of pupil size during the task (e.g., padding with a mirror image of the signal) are not a good estimate of what was happening before the task and in the first few seconds of the first trial. Conversely, the dynamics at the end of the task are likely a better model for what will continue to happen during the next hundreds of milliseconds after the task ends.

We downsampled the pupil time series just for deconvolutions to speed up the computation time. Every eighth sample from the time series was preserved, respecting the Nyquist frequency (this was after the data were low-pass–filtered with a cutoff of 10 Hz). Of the 150 runs of eye data, 7 were removed because of eye tracker malfunction/signal quality.

One notable preprocessing method (*29*) regresses out post-saccade pupil responses. On the basis of our results, we advise against doing this because these "post-ocular events" share variance with the task-evoked pupil response—indeed, they arise from the same system.

Saccade detection and rate estimation

Saccades (including microsaccades) were detected with the method of Engbert and Mergenthaler (*89*). A duration cutoff of 7 ms and a velocity cutoff of 8 times the SD of the horizontal and vertical velocity (degrees/second) were used as inclusion criteria. The distribution of saccade amplitudes is unimodal, which has led to varying operational definitions of a microsaccade (*58*). Our cutoff (which was similar to but slightly more liberal than Engbert and Mergenthaler's) includes small saccades as well as microsaccades. We refer to both of these as "saccades" throughout. We included both, which are known to have different properties (e.g., the former is voluntary, but the latter is involuntary) (*90*, *91*), because previous studies show that a similar constrictory pupil response is observed following either (*28*–*30*).

We quantified the impact of including larger saccades [i.e., greater than 1.5° in amplitude (*91*)] on our results by repeating our model fitting for each observer and excluding any saccades above 1.5° in amplitude from all steps. The stimulus was positioned 5° diagonally from fixation, and the screen edge was 20° horizontally and 15° vertically from the screen edge. Therefore, larger saccades might evoke a pupillary light reflex. The proportion of saccades larger than 1.5° was 3.12% (average across 15 datasets), i.e., quite rare. The median increase in R^2 across observers before versus after, including an upper threshold on saccade amplitude of 1.5° was 0.3% [median absolute deviation (MAD), 0.7%; $N = 15$ datasets]. The median increase in gain was 0.00011 (MAD, 0.00034; $N = 15$ datasets). Thus, the impact of larger saccades on our model fitting was negligible, and we obtained comparable results with or without them for our chosen task.

Estimation of the saccade-locked pupil response

For each run of trials, we estimated the saccade-locked pupil response using deconvolution, a common method for determining the finite impulse response function (IRF) of a linear system. We used an algorithm similar to the one used by Knapen *et al.* (*29*), in which we defined a deconvolution design matrix (i.e., a Toeplitz matrix). The first column contained ones at the onset times of each saccade and zeros everywhere else. Successive columns were shifted by one sample. We estimated the IRF by inverting the linear system, given the entire measured pupil size time course. We assumed a 4-s duration for the saccade-locked pupil response, which was sufficient to capture the whole response according to informal comparisons of different durations and previous studies (*29*).

Linear-nonlinear linking model

We used a linear-nonlinear linking model to predict pupil size from saccade rate and to estimate arousal level. The system's input, the generator function, was assumed to be a time-varying signal corrupted by additive Gaussian noise with SD σ equal to 1. The system has two "channels" (Fig. 1B), one leading from the generator function to saccades (Fig. 2A) and the other leading from the generator function to pupil size (Fig. 2E).

We can express the generator function on a single trial as

$$X_t \sim \mathcal{N}(\mu_t, \sigma) \tag{1}$$

where X_t is the generator function, x represents the possible values it can take, t is time from trial onset to end, dt is determined by the sampling rate of the eye tracker (i.e., dt = 1/500 s for our data), and μ_t is the expected value of the generator function (i.e., across trials).

The channel leading from the generator function to saccades contains two operations, the first a threshold nonlinearity and the second a post-saccade refractory period. The threshold nonlinearity simply subjects the noisy generator function to a fixed threshold, represented in Fig. 1B as a step function. Any time the generator function rises above the threshold is represented by a one or otherwise by a zero in a binary sequence. This binary sequence represents the time course of saccades (and ignores saccade amplitude). The threshold is implemented by mapping trial-averaged saccade rate through a nonlinear function (the inverse cumulative distribution function of a Gaussian with SD σ = 1). The threshold nonlinearity is similar to an inhomogeneous Poisson process (see the "Equivalence of our model with an inhomogeneous Poisson process" section in the Supplementary Materials). The second computation, the post-saccadic refractory period, models the phenomenon of inhibition in saccade generation after a saccade has just occurred as a multiplicative scaling of the entire trial-averaged saccade rate function (see the "Modeling the post-saccade refractory period" section in the Supplementary Materials). We adopted a statistical model called a renewal process to account for the refractory period by starting with a Poisson process and scaling the rate by 1/k—equivalent to preserving every kth saccade (*92*). The Poisson process has an exponential distribution of inter-saccade intervals, but the renewal process has a gamma distribution. This channel of the model yields an (adjusted) trial-averaged saccade rate function. We can therefore write the mapping between saccade rate and the generator function as

$$\lambda_t = \frac{1}{k}\int_h^\infty \mathcal{N}(x; \mu_t, \sigma)\, dx \tag{2}$$

λ_t is saccade rate (a probability between 0 and 1), k is the shape parameter of a gamma distribution fit to the inter-saccade intervals, and h is a fixed threshold (arbitrary, but fixed to 1 in our model fitting and simulations).

The channel leading from the generator function to pupil size contains two operations, a multiplicative gain (i.e., an amplifier) and a low-pass filter (i.e., a temporal integrator). The gain simply scales the noisy generator function by a number g. We assume that gain is constant within a trial and may change across trials but that its expected value across trials is g. The model's prediction of the task-evoked pupil response is equal to the convolution of the linear filter (see the "Estimating the linear filter" section) with the expected value of the mean-subtracted, gain-modulated generator function, plus an additive offset. Mean-subtracting the generator function before the gain corresponds with a particular sort of gain modulation, in which the DC offset of the generator function does not change with gain (like a stereo amplifier might do); only its amplitude (distance from peak to trough) changes. The model fits data equally well without mean-subtracting the generator function, but not mean-subtracting makes three unjustifiable assumptions: (i) that the gain and DC offset of pupil size are strongly (positively) correlated, (ii) that the DC offset is (much) higher than the mean pupil size, and (iii) that the threshold depends on gain. Instead, we simply mean-subtract the generator function and later add to the model prediction an additive offset b, equal to the mean pupil size over a run of trials. This second channel of the model yields a prediction of trial-averaged pupil area (i.e., the task-evoked pupil response). We can write the relation between the task-evoked pupil response and the expected value of the generator function as

$$p_t = l_t * (g\mu_t - \mathbf{E}_t[g\mu_t]) + b \tag{3}$$

where l_t is the linear filter, g is the gain, b is the additive offset, "$*$" denotes (circular) convolution, and $\mathbf{E}$ denotes an expectation. The mean and SD of the gain-modulated generator function are

$$\mathbf{E}_x[gX_t] = g\mu_t \tag{4}$$

$$\sqrt{\mathrm{Var}(gX_t)} = g \tag{5}$$

Therefore, both the amplitude and SD of the generator function scale with gain.

Parameter estimation

We estimated the generator function and gain using trial-averaged pupil and saccade data. We colloquially refer to this estimation method as "ascending one channel of the model and descending the other," because we started with measured saccade rate, estimated the expected value of the generator function, and finally predicted the trial-averaged pupil response, which can be compared to the measured pupil response (Fig. 2). This procedure yielded estimates of the expected value of the generator function and three parameters—gain, offset, and k—of which only gain was fit separately for each run.

To estimate gain, we performed a linear regression between the measured and predicted trial-averaged pupil response. Gain was equal to the beta value of this regression; i.e., it simply scaled the prediction. The predicted pupil response was computed as the convolution of the linear filter and the mean-subtracted and amplified expected value of the generator function, plus an additive offset (mean pupil size over a run) (Eq. 3).

To compute the expected value of the generator function, we mapped the adjusted saccade rate function (i.e., "adjusted": times k, see below) through the inverse function of a cumulative normal distribution. Saccade rate at time t was assumed to be the integral of a Gaussian with SD $\sigma = 1$ beyond a fixed threshold (i.e., from the threshold to infinity). The mean of this Gaussian was the expected value of the generator function at time t. The threshold was always assumed to be 1, but its value is arbitrary as long as it is fixed. We can write the expected value of the generator function μ_t as a function of saccade rate λ_t by inverting Eq. 2

$$\mu_t = h - \Phi^{-1}(1 - k\lambda_t; 0, \sigma) \tag{6}$$

Φ is the cumulative probability function for a Gaussian distribution, and Φ^{-1} is its inverse. We use Eq. 6 to estimate the expected generator function μ_t.

The entire transformation between trial-averaged saccade rate and pupil area can be expressed as

$$p_t = l_t * g[(h - \Phi^{-1}(1 - k\lambda_t; 0, 1)) - \mathbf{E}_t[h - \Phi^{-1}(1 - k\lambda_t; 0, 1)]] + b \tag{7}$$

Note that there is a nonlinearity in between saccade rate and pupil area: the inverse cumulative Gaussian.

We estimated k, the parameter controlling the post-saccadic refractory period, for each observer by fitting a gamma distribution to the empirical distribution of inter-saccadic intervals (for justification, see the "Modeling the post-saccade refractory period" section in the Supplementary Materials). k is the shape parameter of a gamma distribution. For each run, we multiplied the saccade rate function by k to compensate for the reduction in saccades caused by the refractory period, equivalent to preserving only every kth saccade (*92*). The multiplier k on saccade rate scales the input to the nonlinearity Φ, meaning that k influences pupil size in a nonlinear way (Eq. 7). In practice, however, including k in the model had a negligible effect on the model's parameter estimates and goodness of fit (see the Supplementary Materials). Thus, although modeling the refractory period was theoretically justified, it was practically unnecessary for real data. This was true under the assumption of a gamma renewal process or under the assumption of an absolute post-saccadic refractory period of 125 ms (*93*). We confirmed that our model fitting procedure was able to recover the gain and generator function, when specified a priori, with parameter recovery simulations (see the "Parameter recovery: our model" section in the Supplementary Materials).

Estimating the linear filter

The linear filter was estimated by fitting a parametric form to the deconvolved saccade-locked pupil response (Fig. S3A). We optimized for the best combination of parameters (n and $t_{\max}$) that best described the saccade-locked pupil response as a gamma-Erlang function, a common approach in the pupil modeling literature (*22*, *23*, *29*). We used the fminsearch function (Nelder-Mead) in Matlab for the optimization. The equation for the parametric form was

$$l_t = -ft^n e^{-\frac{nt}{t_{\max}}} \tag{8}$$

The initial points for n and $t_{\max}$ were randomly drawn from the intervals (0,20) and (600,4000), respectively. These initial points were chosen on the basis of the parameter range used in previous studies (*22*, *23*, *29*). Maximum function evaluation and maximum iteration both were set to 5×10^4. One set of parameters were fitted for each observer. We set f (the scale factor) such that the parametric form had the same height (minimum) as the measured saccade-locked pupil response, equivalent to multiplying the estimated filter by a normalization factor (height of filter estimate divided by height of saccade-locked pupil response). This meant that estimates of gain were relative to the amplitude of each observer's saccade-evoked pupil response. Therefore, gain estimates were in the same range for all observers, and gain modulation (i.e., with task difficulty) could be compared directly across observers (Fig. 3C). This normalization was a crucial preprocessing step (*88*) because foreshortening error varies for each observer because of the angle of the eye tracker (i.e., adjusted for each observer's height), which affects the range of measured pupil response amplitudes (and thus, gains, also). Note that this is different from divisive baseline correction (*22*, *50*, *94*), in which the pupil size signal is divided by its mean to account for a supposedly nonlinear interaction between the signal's amplitude and additive offset. We did not perform divisive baseline correction on our data because the task-evoked pupil response has been shown to scale linearly with baseline within normal range (i.e., not at the highest or lowest pupil sizes, where there must be compressive nonlinearities due to biomechanical limits) (*50*). We set f based on the minimum of the saccade-locked pupil response because the alternative, regression, caused misestimation of the filter's amplitude and other parameters due to distortion of the saccade-locked response (caused by the threshold nonlinearity). Setting f in the way that we did avoided this, according to our simulations (see below).

The time-to-peak ($t_{\max}$) and width of this filter (n) varied per observer, and these idiosyncrasies were critical for explaining variability in the dynamics of the task-evoked pupil response for some observers (see the "Individual differences in saccade rate and linear filter were sometimes critical to accurately predict pupil size" section in the Supplementary Materials), as Denison *et al.* also suggested (*22*).

We used an (upside-down) gamma-Erlang function as a parametric form for three reasons. First, our assumption was that the linear filter is low-pass and constrictory. Second, a gamma-Erlang function has been shown to model the aggregate input-output relationship of a system with cascades of exponential computations with different time constants (e.g., a multisynaptic neural circuit like the parasympathetic pathway of the pupil) (*23*). Third, parameter recovery simulations revealed that this parametric form eliminates a large amount of bias in the filter estimate relative to a ground truth filter (see the "Justification for linear filter estimation method" section in the Supplementary Materials).

Model comparison

To assess model generalizability, we computed the reduction in R^2 between the in- versus out-of-sample fits as well as ASI and HSI indexes (see Results for details). For our model, we computed the average gain across all easy and hard runs separately for each observer and used these average parameter values for the out-of-sample model predictions. For the consensus model, we did the same but with the best-fit amplitudes for the impulses at trial onset, button press, and for the time-on-task boxcar for the easy and hard data. The duration of the time-on-task boxcar was from trial onset to the mean reaction time. To fit the consensus model, we used the Pupil Response Estimation Toolbox (PRET) toolbox (*22*) to implement a model similar to that proposed by Denison *et al.* (*22*) and others (*17*, *18*, *29*). These models are variants of one another but share the same overarching assumptions: a linear model with impulses near trial events as input and a dilatory linear filter. We did not fit the latency of the impulses from the task events, as Denison and colleagues suggest doing, because it was not standard and would afford the model too much flexibility (also see the "Parameter recovery: Deconvolution method of Wierda and colleagues" section in the Supplementary Materials). When we did fit these latencies, the in-sample fit of the model improved, and the generalizability decreased, consistent with the idea that it made the overfitting worse.

For the consensus model only, the computed ASI and HSI indexes were sometimes extremely large or small because one of the best-fit amplitudes would go to nearly zero on one dataset and would be larger than zero for the other dataset. This suggested that the three parameters traded off in the consensus model, again suggesting that it overfits.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abi9979

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. I. E. Loewenfeld, O. E. Lowenstein, *The Pupil: Anatomy, Physiology, and Clinical Applications* (Butterworth-Heinemann, ed. 2, 1999).
2. R. S. Larsen, J. Waters, Neuromodulatory correlates of pupil dilation. *Front. Neural Circuits* **12**, 21 (2018).
3. D. H. McDougal, P. D. Gamlin, Autonomic control of the eye. *Compr. Physiol.* **5**, 439–473 (2014).
4. P. R. K. Turnbull, N. Irani, N. Lim, J. R. Phillips, Origins of pupillary hippus in the autonomic nervous system. *Investig. Ophthalmol. Vis. Sci.* **58**, 197–203 (2017).
5. O. Lowenstein, R. Feinberg, I. E. Loewenfeld, Pupillary movements during acute and chronic fatigue a new test for the objective evaluation of tiredness. *Invest. Ophthalmol. Vis. Sci.* **2**, 138–157 (1963).
6. E. Hess, J. M. Polt, Pupil size in relation to mental activity during simple problem-solving. *Science* **143**, 1190–1192 (1964).
7. D. Kahneman, J. Beatty, Pupil diameter and load on memory. *Science* **154**, 1583–1585 (1966).
8. G. Aston-Jones, J. D. Cohen, An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
9. C. Willems, J. Herdzin, S. Martens, Individual differences in temporal selective attention as reflected in pupil dilation. *PLOS ONE* **10**, e0145056 (2015).
10. C.-A. Wang, T. Baird, J. Huang, J. D. Coutinho, D. C. Brien, D. P. Munoz, Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. *Front. Neurol.* **9**, 1029–1042 (2018).
11. J. Reimer, M. J. McGinley, Y. Liu, C. Rodenkirch, Q. Wang, D. A. McCormick, A. S. Tolias, Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **7**, 13289 (2016).
12. S. Joshi, Y. Li, R. M. Kalwani, J. I. Gold, Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* **89**, 221–234 (2016).
13. F. Cazettes, D. Reato, J. P. Morais, A. Renart, Z. Mainen, Phasic activation of dorsal raphe serotonergic neurons increases pupil size. *Curr. Biol.* **31**, 192–197.e4 (2021).
14. M. J. McGinley, S. V. David, D. A. McCormick, Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron* **87**, 179–192 (2015).
15. S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, A. K. Churchland, Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).
16. S. Joshi, J. Gold, Pupil size as a window on neural substrates of cognition. *Trends Cogn. Sci.* **24**, 466–480 (2020).
17. J. W. de Gee, T. Knapen, T. H. Donner, Decision-related pupil dilation reflects upcoming choice and individual bias. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 618–625 (2014).
18. P. R. Murphy, E. Boonstra, S. Nieuwenhuis, Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nat. Commun.* **7**, 13526 (2016).
19. J. W. de Gee, O. Colizoli, N. A. Kloosterman, T. Knapen, S. Nieuwenhuis, T. H. Donner, Dynamic modulation of decision biases by brainstem arousal systems. *eLife* **6**, e23232 (2017).
20. C. W. Korn, M. Staib, A. Tzovara, G. Castegnetti, D. R. Bach, A pupil size response model to assess fear learning. *Psychophysiology* **54**, 330–343 (2017).
21. J. C. V. Slooten, S. Jahfari, T. Knapen, J. Theeuwes, How pupil responses track value-based decision-making during and after reinforcement learning. *PLoS Comput. Biol.* **14**, e1007031 (2018).
22. R. N. Denison, J. A. Parker, M. Carrasco, Modeling pupil responses to rapid sequential events. *Behav. Res. Methods* **52**, 1991–2007 (2020).
23. B. Hoeks, W. J. M. Levelt, Pupillary dilation as a measure of attention: A quantitative system analysis. *Behav. Res. Methods Instrum. Comput.* **25**, 16–26 (1993).
24. S. M. Wierda, H. van Rijn, N. A. Taatgen, S. Martens, Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8456–8460 (2012).

25. C.-A. Wang, S. E. Boehnke, B. J. White, D. P. Munoz, Microstimulation of the monkey superior colliculus induces pupil dilation without evoking saccades. *J. Neurosci.* **32**, 3629–3636 (2012).
26. C.-A. Wang, D. P. Munoz, A circuit for pupil orienting responses: Implications for cognitive modulation of pupil size. *Curr. Opin. Neurobiol.* **33**, 134–140 (2015).
27. C.-A. Wang, D. P. Munoz, Coordination of pupil and saccade responses by the superior colliculus. *J. Cogn. Neurosci.* **33**, 919–932 (2021).
28. B. Zuber, L. Stark, M. Lorber, Saccadic suppression of the pupillary light reflex. *Exp. Neurol.* **14**, 351–370 (1966).
29. T. Knapen, J. W. de Gee, J. Brascamp, S. Nuiten, S. Hoppenbrouwers, J. Theeuwes, Cognitive and ocular factors jointly determine pupil responses under equiluminance. *PLOS ONE* **11**, e0155574 (2016).
30. A. Benedetto, P. Binda, Dissociable saccadic suppression of pupillary and perceptual responses to light. *J. Neurophysiol.* **115**, 1243–1251 (2016).
31. S. Joshi, Pupillometry: Arousal state or state of mind? *Curr. Biol.* **31**, R32–R34 (2021).
32. I. E. Loewenfeld, Mechanisms of reflex dilatation of the pupil. *Doc. Ophthalmol.* **12**, 185–448 (1958).
33. M. A. Phillips, E. Szabadi, C. M. Bradshaw, Comparison of the effects of clonidine and yohimbine on spontaneous pupillary fluctuations in healthy human volunteers. *Psychopharmacology (Berl)* **150**, 85–89 (2000).
34. M. A. Phillips, E. Szabadi, C. M. Bradshaw, Comparison of the effects of clonidine and yohimbine on pupillary diameter at different illumination levels. *Br. J. Clin. Pharmacol.* **50**, 65–68 (2000).
35. S. Steinhauer, R. Condray, M. Pless, Pharmacological isolation of cognitive components influencing the pupillary light reflex. *J. Ophthalmol.* **2015**, 179542 (2015).
36. M. P. Bokoch, M. Behrends, A. Neice, M. Larson, Fentanyl, an agonist at the mu opioid receptor, depresses pupillary unrest. *Auton. Neurosci.* **189**, 68–74 (2015).
37. C. Strauch, L. Greiter, A. Huckauf, Pupil dilation but not microsaccade rate robustly reveals decision formation. *Sci. Rep.* **8**, 13165 (2018).
38. R. N. Denison, S. Yuval-Greenberg, M. Carrasco, Directing voluntary temporal attention increases fixational stability. *J. Neurosci.* **39**, 353–363 (2019).
39. M. Rolfs, R. Kliegl, R. Engbert, Toward a model of microsaccade generation: The case of microsaccadic inhibition. *J. Vis.* **8**, 5.1–5.523 (2008).
40. S. Martinez-Conde, S. L. Macknik, Unchanging visions: The effects and limitations of ocular stillness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160204 (2017).
41. D. Abeles, R. Amit, N. Tal-Perry, M. Carrasco, S. Yuval-Greenberg, Oculomotor inhibition precedes temporally expected auditory targets. *Nat. Commun.* **11**, 3524 (2020).
42. A. L. White, M. Rolfs, Oculomotor inhibition covaries with conscious detection. *J. Neurophysiol.* **116**, 1507–1521 (2016).
43. S. Stanten, L. Stark, A statistical analysis of pupil noise. *I.E.E.E. Trans. Biomed. Eng.* **13**, 140–152 (1966).
44. J. Krueger, A. A. Disney, Structure and function of dual-source cholinergic modulation in early vision. *J. Comp. Neurol.* **527**, 738–750 (2019).
45. A. E. Urai, A. Braun, T. H. Donner, Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 1–11 (2017).
46. O. Colizoli, J. W. de Gee, A. E. Urai, T. Donner, Task-evoked pupil responses reflect internal belief states. *Sci. Rep.* **8**, 1–13 (2018).
47. R. B. Ebitz, T. Moore, Both a gauge and a filter: Cognitive modulations of pupil size. *Front. Neurol.* **9**, 1190–1203 (2019).
48. J. Smith, L. Y. Ichinose, G. A. Masek, T. Watanabe, L. Stark, Midbrain single units correlating with pupil response to light. *Science* **162**, 1302–1303 (1968).
49. R. Suzuki, H. Yoshino, S. Kobayashi, Different time courses of bovine iris sphincter and dilator muscles after stimulation. *Ophthalmic Res.* **19**, 344–350 (1987).
50. J. Reilly, A. Kelly, S. Kim, S. Jett, B. M. Zuckerman, The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behav. Res. Methods* **51**, 865–878 (2019).
51. C. W. Korn, D. R. Bach, A solid frame for the window on cognition: Modeling event-related pupil responses. *J. Vis.* **16**, 28 (2016).
52. M. Megemont, J. McBurney-Lin, H. Yang, Pupil diameter is not an accurate real-time readout of locus coeruleus activity. *eLife* **11**, e70510 (2022).
53. G. M. Boynton, S. A. Engel, G. H. Glover, D. J. Heeger, Linear systems analysis of functional magnetic resonance imaging in human v1. *J. Neurosci.* **16**, 4207–4221 (1996).
54. K. N. Kay, J. Winawer, A. Rokem, A. Mezer, B. Wandell, A two-stage cascade model of bold responses in human visual cortex. *PLoS Comput. Biol.* **9**, e1003079 (2013).
55. M. Carandini, D. Heeger, Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
56. J. S. Antrobus, Information theory and stimulus-independent thought. *Br. J. Psychol.* **59**, 423–430 (1968).
57. L. M. Giambra, A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Conscious. Cogn.* **4**, 1–21 (1995).
58. S. Martinez-Conde, J. Otero-Millan, S. L. Macknik, The impact of microsaccades on vision: Towards a unified theory of saccadic function. *Nat. Rev. Neurosci.* **14**, 83–96 (2013).
59. J.-T. Chen, R. Yep, Y.-F. Hsu, Y.-G. Cherng, C.-A. Wang, Investigating arousal, saccade preparation, and global luminance effects on microsaccade behavior. *Front. Hum. Neurosci.* **15**, 602835 (2021).
60. J. D. Hunter, J. Milton, H. Luedtke, B. Wilhelm, H. Wilhelm, Spontaneous fluctuations in pupil size are not triggered by lens accommodation. *Vision Res.* **40**, 567–573 (2000).
61. Z. M. Hafed, L. Goffart, R. J. Krauzlis, A neural mechanism for microsaccade generation in the primate superior colliculus. *Science* **323**, 940–943 (2009).
62. S. Nummela, R. Krauzlis, Inactivation of primate superior colliculus biases target choice for smooth pursuit, saccades, and button press responses. *J. Neurophysiol.* **104**, 1538–1548 (2010).
63. B. White, J. Y. Kan, R. Levy, L. Itti, D. Munoz, Superior colliculus encodes visual saliency before the primary visual cortex. *Proc. Natl. Acad. Sci.* **114**, 9451–9456 (2017).
64. K. H. Lee, A. Tran, Z. Turan, M. Meister, The sifting of visual information in the superior colliculus. *eLife* **9**, 1–23 (2020).
65. L. Wang, K. McAlonan, S. Goldstein, C. Gerfen, R. Krauzlis, A causal role for mouse superior colliculus in visual perceptual decision-making. *J. Neurosci.* **40**, 3768–3782 (2020).
66. D. P. Hanes, J. Schall, Neural control of voluntary movement initiation. *Science* **274**, 427–430 (1996).
67. J. Schall, B. A. Purcell, R. Heitz, G. Logan, T. Palmeri, Neural mechanisms of saccade target selection: Gated accumulator model of the visual-motor cascade. *Eur. J. Neurosci.* **33**, 1991–2002 (2011).
68. L. Breen, R. Burde, A. Loewy, Brainstem connections to the edinger-westphal nucleus of the cat: A retrograde tracer study. *Brain Res.* **261**, 303–306 (1983).
69. S. Nieuwenhuis, E. J. De Geus, G. Aston-Jones, The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology* **48**, 162–175 (2011).
70. T. Kozicz, J. Bittencourt, P. May, A. Reiner, P. D. Gamlin, M. Palkovits, A. Horn, C. A. Toledo, A. Ryabinin, The Edinger-Westphal nucleus: A historical, structural, and functional perspective on a dichotomous terminology. *J. Comp. Neurol.* **519**, 1413–1434 (2011).
71. R. Warwick, The ocular parasympathetic nerve supply and its mesencephalic sources. *J. Anat.* **88**, 71–93 (1954).
72. E. Bruinstroop, G. Cano, V. VanderHorst, J. C. Cavalcante, J. R. Wirth, M. Sena-Esteves, C. Saper, Spinal projections of the A5, A6 (locus coeruleus), and A7 noradrenergic cell groups in rats. *J. Comp. Neurol.* **520**, 1985–2001 (2012).
73. S. Mathôt, Pupillometry: Psychology, physiology, and function. *J. Cogn.* **1**, 16 (2018).
74. S. R. Steinhauer, G. J. Siegle, R. Condray, M. Pless, Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *Int. J. Psychophysiol.* **52**, 77–86 (2004).
75. O. Lowenstein, Mutual role of sympathetic and parasympathetic in shaping of the pupillary reflex to light. *Arch. Neurol. Psychiatry* **64**, 341–377 (1950).
76. P. Heller, F. Perry, D. L. Jewett, J. Levine, Autonomic components of the human pupillary light reflex. *Invest. Ophthalmol. Vis. Sci.* **31**, 156–162 (1990).
77. R. B. Ebitz, T. Moore, Selective modulation of the pupil light reflex by microstimulation of prefrontal cortex. *J. Neurosci.* **37**, 5008–5018 (2017).
78. L. Stark, Stability, oscillations, and noise in the human pupil servomechanism. *Proc. IRE* **47**, 1925–1939 (1959).
79. W. C. Krenz, L. Stark, Systems model for pupil size effect. *Biol. Cybern.* **51**, 391–397 (1985).
80. S. Usui, L. Stark, A model for nonlinear stochastic behavior of the pupil. *Biol. Cybern.* **45**, 13–21 (1982).
81. J. Drewes, W. Zhu, Y. Hu, X. Hu, Smaller is better: Drift in gaze measurements due to pupil dynamics. *PLOS ONE* **9**, e111197 (2014).
82. K. W. Choe, R. Blake, S. hun Lee, Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. *Vision Res.* **118**, 48–59 (2016).
83. S. Mathôt, J.-B. Melmi, E. Castet, J. M. Abdullah, Intrasaccadic perception triggers pupillary constriction. *PeerJ* **3**, e1150 (2015).
84. J. Slooter, D. van Norren, Visual acuity measured with pupil responses to checkerboard stimuli. *Invest. Ophthalmol. Vis. Sci.* **19**, 105–108 (1980).
85. S. Mathôt, Tuning the senses: How the pupil shapes vision at the earliest stage. *Annu. Rev. Vis. Sci.* **6**, 433–451 (2020).
86. L. Kooijman, D. Dodou, S. Jansen, T. Themans, J. Russell, S. Petermeijer, J. Doorman, J. Habl, D. Neubert, M. Vos, J. de Winter, Is accommodation a confounder in pupillometry research? *Biol. Psychol.* **160**, 1–13 (2021).
87. D. M. Dacey, H. Liao, B. B. Peterson, F. R. Robinson, V. Smith, J. Pokorny, K. Yau, P. D. Gamlin, Melanopsin-expressing ganglion cells in primate retina signal colour and irradiance and project to the lgn. *Nature* **433**, 749–754 (2005).
88. T. R. Hayes, A. A. Petrov, Mapping and correcting the influence of gaze position on pupil size measurements. *Behav. Res. Methods* **48**, 510–527 (2016).
89. R. Engbert, K. Mergenthaler, Microsaccades are triggered by low retinal image slip. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7192–7197 (2006).

90. K. Mergenthaler, R. Engbert, Microsaccades are different from saccades in scene perception. *Exp. Brain Res.* **203**, 753–757 (2010).
91. P. Sinn, R. Engbert, Small saccades versus microsaccades: Experimental distinction and model-based unification. *Vision Res.* **118**, 132–143 (2016).
92. J. Victor, S. Nirenberg, Spike trains as event sequences: fundamental implications, in *Spike Timing: Mechanisms and Function* (Boca Raton: Taylor & Francis/CRC Press, 2013), pp. 3–32.
93. R. Amit, D. Abeles, S. Yuval-Greenberg, Transient and sustained effects of stimulus properties on the generation of microsaccades. *J. Vis.* **19**, 6 (2019).
94. S. Mathôt, J. Fabius, E. V. Heusden, S. V. der Stigchel, Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav. Res. Methods* **50**, 94–106 (2018).
95. D. McKeegan, Spontaneous and odour evoked activity in single avian olfactory bulb neurones. *Brain Res.* **929**, 48–58 (2002).
96. G. Maimon, J. Assad, Beyond poisson: Increased spike-time regularity across primate parietal cortex. *Neuron* **62**, 426–440 (2009).
97. B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, J.-S. S. White, Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.* **24**, 127–135 (2009).

Acknowledgments: We thank M. Landy, R. Denison, E. Simoncelli, F. Rieke, and N. Steinmetz for comments on the research and E. Merriam and N. Arfaei for help with early versions of the experiments. **Funding:** This research was supported by a National Eye Institute grant (R01-EY025330) to D.J.H. and a National Defense Science and Engineering Graduate fellowship to C.S.B. **Author contributions:** Conceptualization: S.M., C.S.B., and D.J.H. Methodology: C.S.B., S.M., and D.J.H. Software: C.S.B. and S.M. Formal analysis: C.S.B. and S.M. Investigation: C.S.B. and S.M. Writing—original draft: C.S.B. and S.M. Writing—review and editing: C.S.B., S.M., and D.J.H. Visualization: C.S.B. Funding acquisition: D.J.H. Supervision: D.J.H. **Model development:** S.M., C.S.B., and D.J.H. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Our code toolbox is available at https://github.com/csb0/PCDM and our experimental data and code are available at https://archive.nyu.edu/handle/2451/63809.

Submitted 14 April 2021
Accepted 4 March 2022
Published 20 April 2022
10.1126/sciadv.abi9979