Supplementary Methods

Goodness-of-Fit.
In order to have an independent estimate of the acceptability of our linear regression model, we calculated the goodness-of-fit probability *Q* (Press et al., 1997) for every neuron and every time bin. To estimate *Q* we need the quantity

$$\chi^2(t) \equiv \sum_{i=1}^{N_T} \left(\frac{r_i(t) - \hat{r}_i(t)}{\sigma_i} \right)^2$$

where the index *i* sums over trials, $r_i(t)$ is the value of the PSTH in trial *i*, $r_i(t)$ is the estimated PSTH at the (f1,f2) pair of trial *i* from the regression model, σ_i is an (independent) estimate of the measurement error of the PSTH.

The probability *Q* estimates how likely it is that the sum χ^2 would be exceeded by chance. If the measurement errors are normally distributed, then the values of χ^2 follow a chi-square distribution, and the the value of Q can be calculated from an upper incomplete gamma function

$$Q(a,x) \equiv \frac{1}{\Gamma(a)} \int_x^\infty e^{-t} t^{a-1} dt$$

where the arguments to *Q* are $a = \chi^2/2$ and $x = \nu/2$ (where ν is, as before, the degrees of freedom less the number of parameters). To estimate the measurement errors σ_i, we calculated the standard deviation of the PSTHs across trials for each value of f1 for model (1) and for each (f1,f2) pair for model (2). In cases where only a single trial was encountered for a particular f1 or (f1,f2) pair, we removed that point from the evaluation of *Q*. Also, because PSTHs can be at minimum zero, they tend to violate the assumption of normality when PSTH values are small. Although the assumption of normality is not a strong constraint in the calculation of *Q* (Press et al., 1997), it is highly sensitive to incorrect estimates of the measurement errors σ_i. Therefore, we imposed a minimum value for σ_i to be 1; all values below the minimum were set to the minimum.

Nonlinear regression.
We performed a nonlinear regression of firing rates to stimulus values f1 and f2 using the function nlinfit provided in Matlab (Mathworks Inc.). We used the hyperbolic tangent as the sigmoidal curve:

$$r(t) = \hat{b}_0(t) + \hat{b}_4(t) \tanh(\hat{b}_1(t) f_1 + \hat{b}_2(t) f_2), \quad \text{(S1)}$$

where r(t) is the value of the PSTH at time t, b0-hat is the baseline firing rate, b4-hat (>0) is the amplitude, and b1-hat and b2-hat define the dependence of f1 and f2 respectively.

If the stimulus sets sampled f1 and f2 values uniformly, the direction of the [b1, b2] vector obtained from a linear fit (eqn. 2) and the vector obtained from the linear-nonlinear cascade fit (eqn. S1) would be the same. The neural data used here was obtained with stimulus sets that had a finite number of f1 and f2 values. Nevertheless, in Supplementary Figure **S3** we show that the difference in the direction of the [b1, b2] vector for the two fits is very small. For simplicity and for consistency in fitting all neurons in the same way, results in the main text are reported using linear fits. None of the qualitative conclusions change if we use the linear-nonlinear fit of eqn. S1.

Angle between linear and nonlinear regression.

The angle $\phi(t)$ between the linear regression coefficients ($b_1(t)$,b_2(t)) from eqn.2 and the nonlinear regression coefficients $(\hat{b}_1(t), \hat{b}_2(t))$ from eqn. S1 were calculated using the dot product between the vectors formed by the two coefficients:

$$\phi(t) = \cos^{-1}\left(\frac{b_1(t)\hat{b}_1(t) + b_2(t)\hat{b}_2(t)}{\sqrt{(b_1(t)^2 + b_2(t)^2)(\hat{b}_1(t)^2 + \hat{b}_2(t)^2)}}\right). \quad \text{(S2)}$$

Supplementary Figures

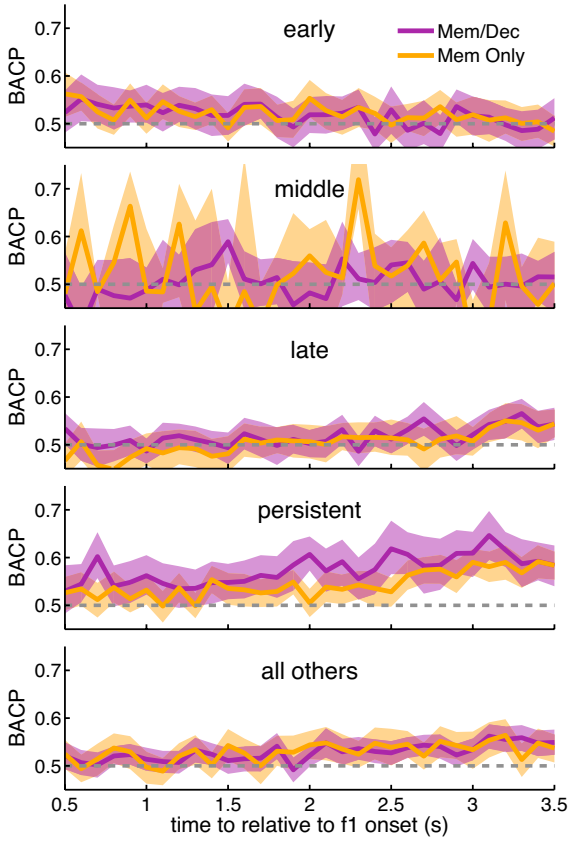

Figure S1. Average choice probability between memory only (Mem Only) cells and memory and decision cells (Mem/Dec) shows that memory activity of Mem Only cells do not correlate better with behavior. Categorization of memory cells and memory and decision cells follow the same procedure used in the main text (see Categorization in Materials and Methods). Populations of cells are further segregated by the time of their significant memory activity as labeled in each graph. Early: first third of delay period only. Middle: second third of delay only. Late: final third of delay period. Persistent: significant in at least 5 of 6 500 ms segments of the delay. All others: all cells not conforming to the above but showing significant memory activity.

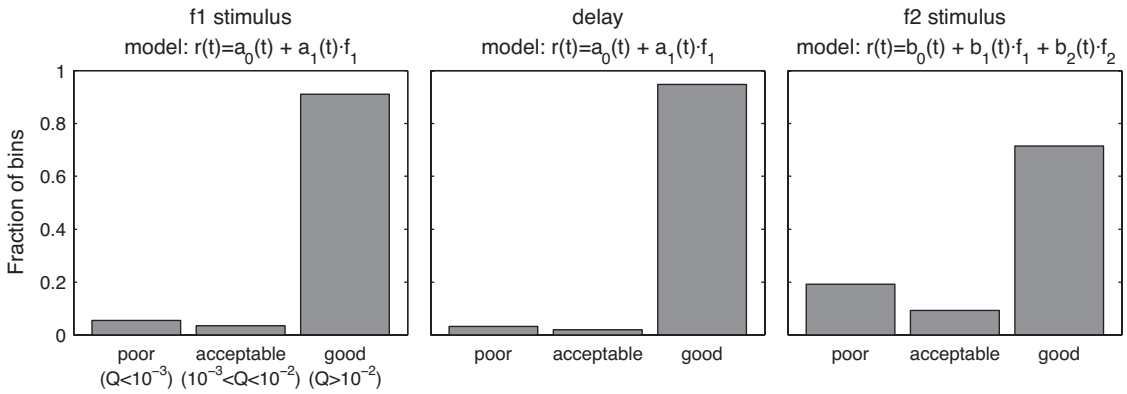

Figure S2. Distribution of goodness-of-fit Q-values for each epoch and corresponding regression model shows linear model is a good fit to the data. For each histogram, all Q-values from time bins used in categorization (see Methods in main text) were calculated. Models are typically accepted with Q>0.001 and are considered good with Q>0.01 (Press et al., 1997). Therefore, Q-values were counted at three levels: poor (Q<0.001), acceptable (0.001<Q<0.01), and good (Q>0.01). The total fraction of Q-values falling into each level for each task epoch are shown in the three panels above: first stimulus period (left), delay period (middle), second stimulus period (right). In all three task epochs, a linear regression model produces a majority of good fits. Calculation of Q is sensitive to an estimate of the measurement error of the PSTH. For the f2 stimulus model, it is more difficult to estimate the measurement error since the number of trials for each (f1,f2) pair is smaller than in the models used during the f1 stimulus and delay periods, where trials can be grouped by f1 only; therefore, the Q-values during the f2 stimulus tend to be worse than in the f1 stimulus and delay periods. In general, using a sigmoid for the regression model should produce better fits as measured by Q, however with the added disadvantages of adding parameters and complicating the analysis of significance. The Q-values shown here demonstrate that a linear model provides a good fit to the data.

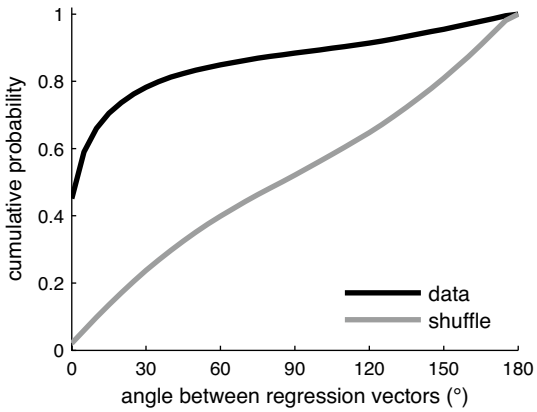

Figure S3. Angular difference between linear regressor and linear-sigmoidal regressor is typically very small. The figure shows the cumulative distributions of angle between linear and nonlinear regression coefficients for data (black) and shuffle (gray) for all time points. Nearly 50% of angles in the data are negligible, and nearly 80% of angles are within 30 degrees; shuffled data is distributed more uniformly across all angles up to 180 degrees. Angles were calculated using eqn. S2 as described in Supplemental Methods. For shuffle data, cell labels for nonlinear regressions were shuffled pseudorandomly while time labels were kept intact.

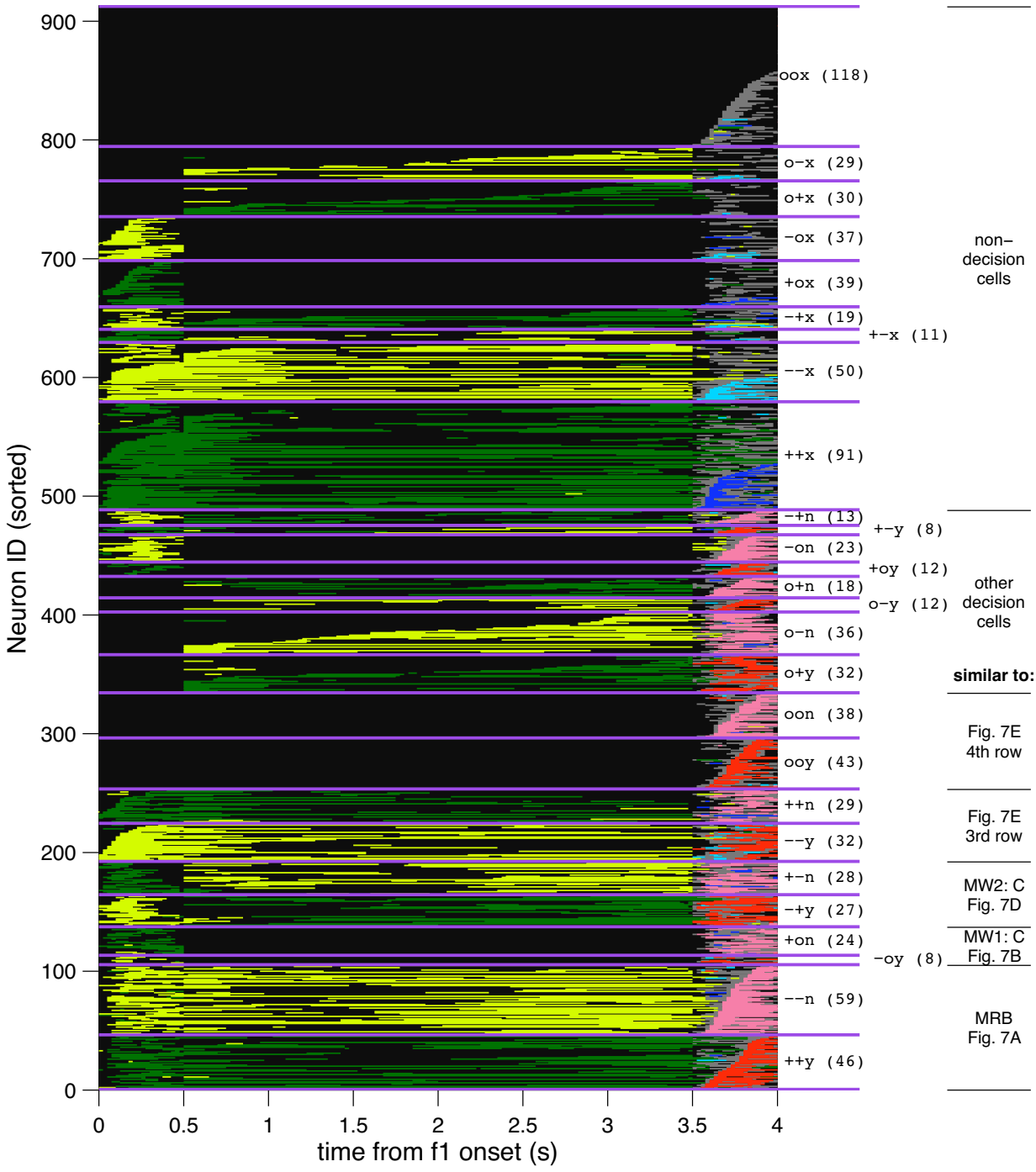
Neuron ID (sorted)
time from f1 onset (s)
0
0.5
1
1.5
2
2.5
3
3.5
4
100
200
300
400
500
600
700
800
900
oox (118)
o−x (29)
o+x (30)
−ox (37)
+ox (39)
−+x (19)
+−x (11)
−−x (50)
++x (91)
−+n (13)
+−y (8)
−on (23)
+oy (12)
o+n (18)
o−y (12)
o−n (36)
o+y (32)
oon (38)
ooy (43)
++n (29)
−−y (32)
+−n (28)
−+y (27)
+on (24)
−oy (8)
−−n (59)
++y (46)
non–decision cells
other decision cells
similar to:
Fig. 7E 4th row
Fig. 7E 3rd row
MW2: C Fig. 7D
MW1: C Fig. 7B
MRB Fig. 7A

Figure S4. Cell encodings over time for the entire population. Data shown here is the same as in Fig. 3 of the main text, however, neuron IDs have been resorted according to the simple categorization scheme described in the main text. Purple lines delineate groups of neurons that belong to the same combination of encoding types. The three character codes to the right indicate encoding type for each group and the numbers in parentheses indicate the number of cells for that group. To decode the encoding type for each group from the three letters: the first letter represents the f1-period, the second letter represents the delay, the third letter represents the f2-period. In the f1-period and the delay, "-" indicates negative f1-encoding, "+" indicates positive f1-encoding, and "o" indicates non-f1-encoding. During the f2-period, "y" indicates "yes"-decision encoding, "n" indicates "no" -decision encoding, and "x" indicates non-decision-encoding (though the cell may encode another variable such as f1 or f2). As an example, a cell in the group "o+y" does not encode f1 during the f1-period, encodes positive f1 during the delay, and encodes the "yes"-decision during the f2-period. On the far, right groups are labeled according to their similarity to the predictions of the theoretical models described in Fig. 7 of the main text. Table S2, below, summarizes the number of cells in each group shown here.

	yes	no	yes/no	+f1	-f1	+f2	-f2	o	sum
++	20	6	0	24	1	24	0	4	**79**
+-	2	8	0	2	0	4	0	0	**16**
+±	3	3	0	1	0	0	0	0	**7**
+o	3	6	0	2	0	9	1	3	**24**
-+	11	5	0	3	3	2	4	2	**30**
--	11	27	0	0	7	0	21	9	**75**
-±	0	0	0	0	0	0	1	0	**1**
-o	3	16	0	0	5	2	6	8	**40**
±+	1	2	0	0	1	0	0	0	**4**
±-	1	3	1	0	0	0	0	0	**5**
±±	0	0	0	0	0	0	0	0	**0**
±o	0	2	0	0	0	1	0	0	**3**
o+	24	17	0	4	0	5	3	7	**60**
o-	8	33	0	0	3	3	5	4	**56**
o±	1	1	0	0	0	0	0	1	**3**
oo	41	36	1	4	3	5	3	72	**165**
sum	**129**	**165**	**2**	**40**	**23**	**55**	**44**	**110**	**568**

Table S1. Number of cells fitting into most of the possible encoding category combinations. Cells were categorized in each epoch using criteria described in the main text. For the f1-stimulus period and delay period, there are four category types: positive f1-encoding (+), negative f1-encoding (-), ambiguous sign f1-encoding (±), and non-encoding (o). For the f2-stimulus period, there are ten category types: "yes"-decision encoding (yes), "no"-decision encoding (no), ambiguous decision encoding (yes/no); +,-, ± f1-encoding (+f1,-f1, ±f1); +,-, ± f2-encoding (+f2,-f2, ±f2); and non-encoding (o). The 16 combinations of f1-period and delay period encodings are placed along the rows. The f2-stimulus encodings types are shown along the columns, but only eight are shown; no cells were found to belong to ±f1 or ±f2. The table shows the number of cells that fit into each encoding type combination along with the sums for each column and row. The total number of cells categorized is shown in the lower right corner. Cells that are not categorized here were ambiguous in their encodings, i.e. they showed too much significant regression to be deemed non-encoding but too little significant regression to be deemed fully encoding in at least one epoch. Entries in green show numbers reported in Fig. 7 except for those in Fig. 7C which are not reported here.