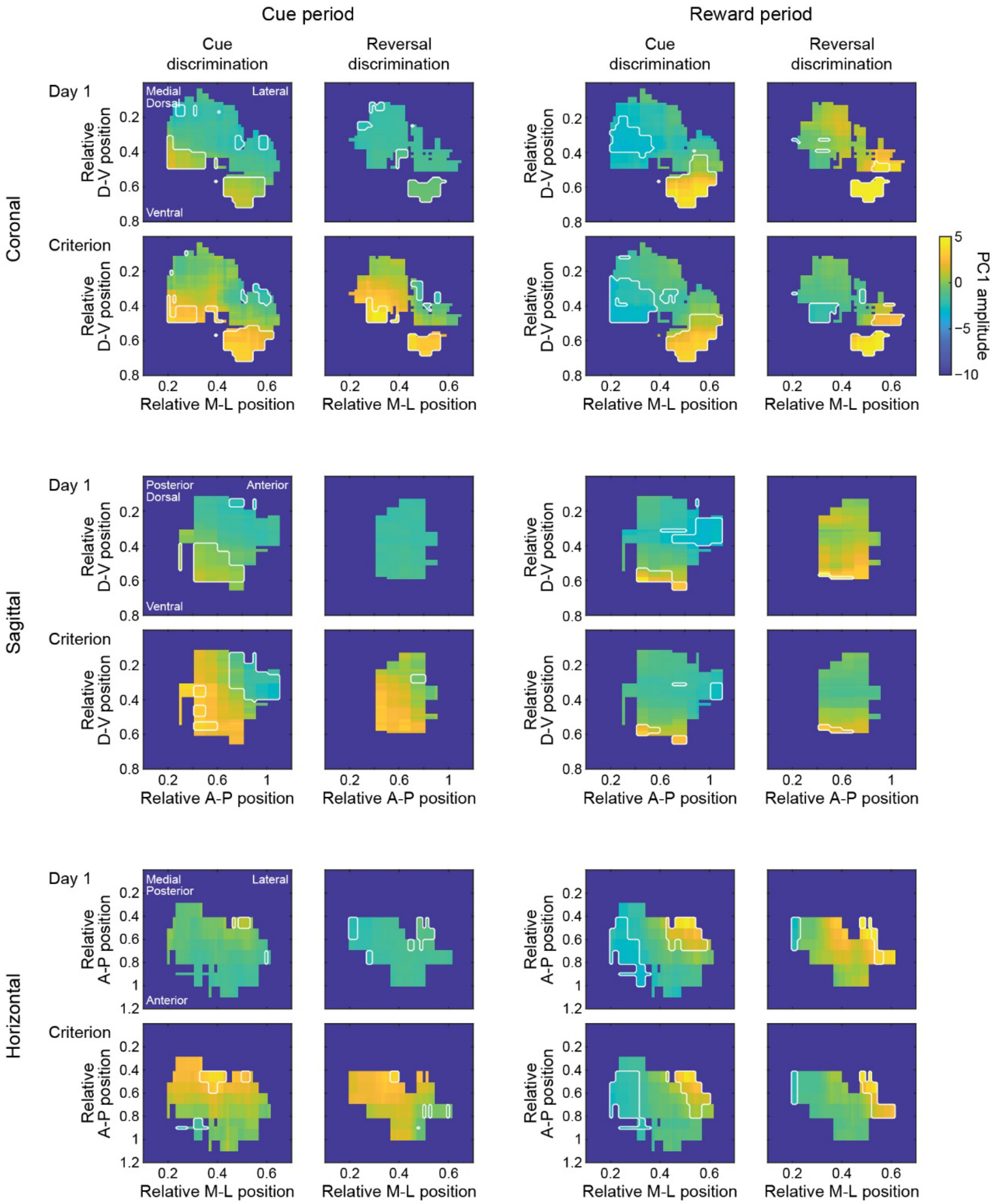
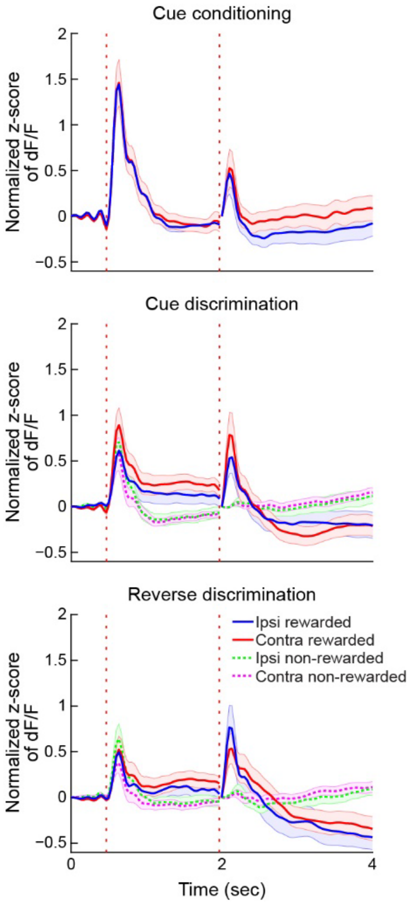
Supplementary Figures

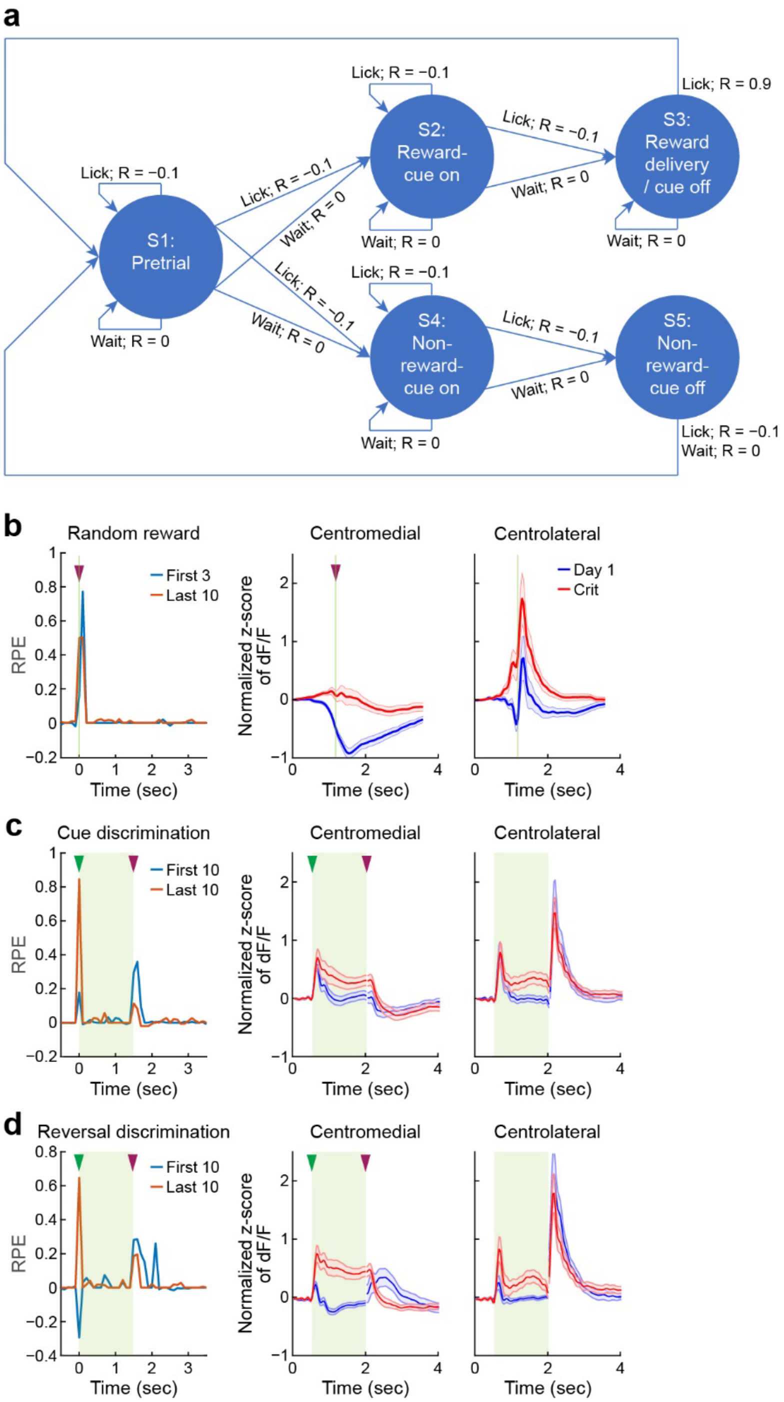


Supplementary Fig. 1. **Anatomical distribution of PC1 amplitude for the first (Day 1) and last (Criterion) sessions in cue discrimination and cue reversal training, shown in all three orthographic projection planes.** For this analysis, the waveforms from all four session sets (cue discrimination Day 1, cue discrimination Criterion, cue reversal Day 1, cue reversal Criterion) were analyzed together in a single PCA run to ensure that the PC amplitudes were directly comparable. The first 3 PC waveforms were similar to those observed when analyzing only the Cue

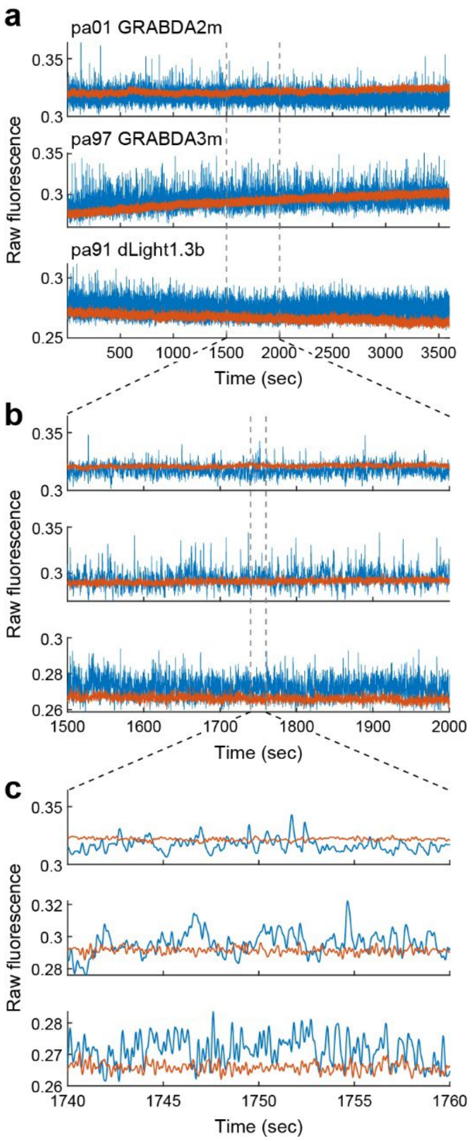
Discrimination − Criterion waveforms (**Fig. 6**), as were the qualitative patterns of the PC1 anatomical maps (with slight differences in amplitude values). A reduction in reward-related PC1 amplitude with learning was observed only in the dorsal-most sites and exclusively during reversal training.



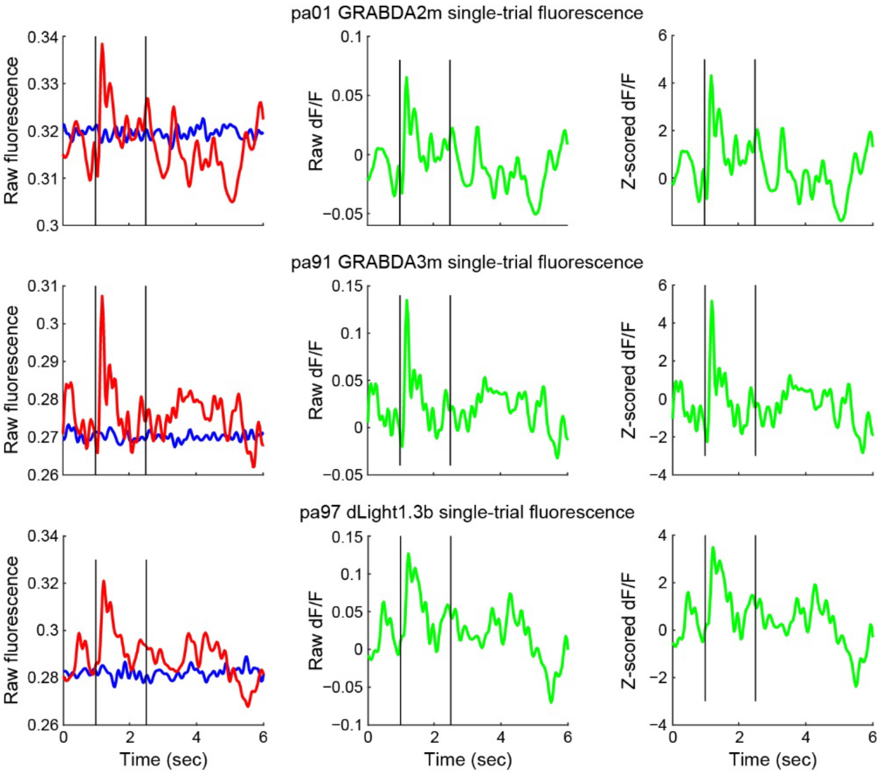
Supplementary Fig. 2. Statistical equivalence of dopamine signals (mean ± 2 SEM) from criterion sessions for cues presented ipsilateral and contralateral to the recording probe. Top: Cue conditioning task. Middle: Cue discrimination task. Bottom: Reversal discrimination task.



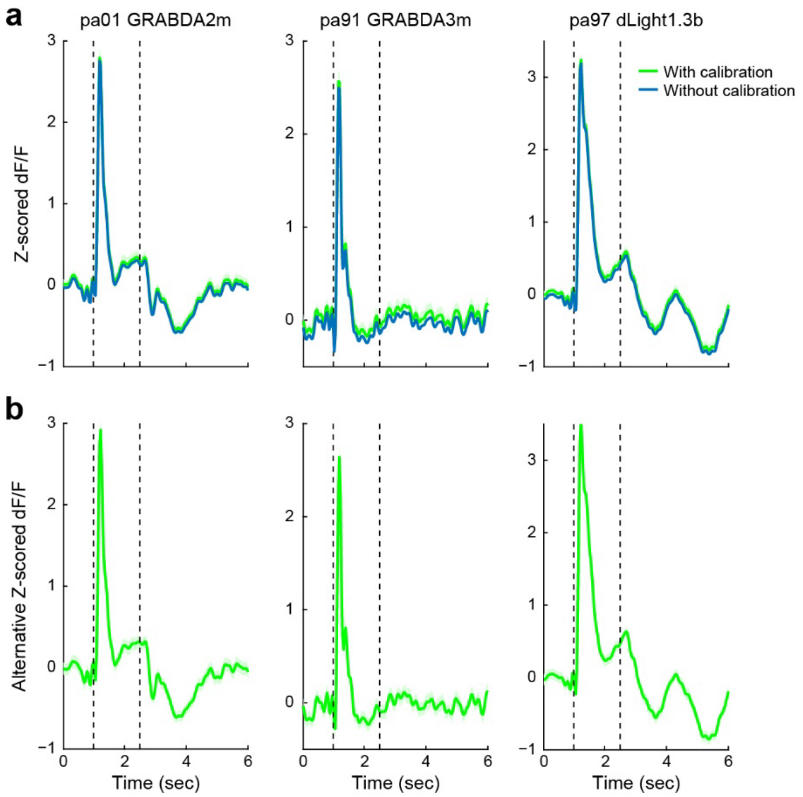
Supplementary Fig. 3. **Comparison of model RPE results with experimental results with dopamine release responses.** See Supplementary Text. **a** State transition diagram for the Q-learning model. S1 to S5 represent model states. Each arrow exiting from a state is annotated with the associated action and its net reward value. **b-d** First column shows trial-averaged RPE values from the Q-learning model; second and third columns are reproduced from main **Fig. 2a** for convenience of comparison. **b** Random reward responses (mean ± 2 SEM). The model was trained solely on the random reward task. Model trials were aligned on reward delivery (time 0) and averaged for comparison to the trial-averaged dopamine responses. First licks typically occurred on the next time step. Blue: first 3 trials; red: last 10 trials. **c** Cue discrimination training. Training began directly with the cue discrimination task. Model trials were aligned on cue onset (time 0). Cue termination and reward delivery occurred 1.5 sec after cue onset. Blue: first 10 trials; red: last 10 trials. Note that the discrimination task plots copied from **Fig. 2a** show cue onset at time 0.5. **d** Reversal discrimination training, aligned as in b. Reversal training followed cue discrimination training to illustrate reversal dynamics. Blue: first 10 trials; red: last 10 trials.



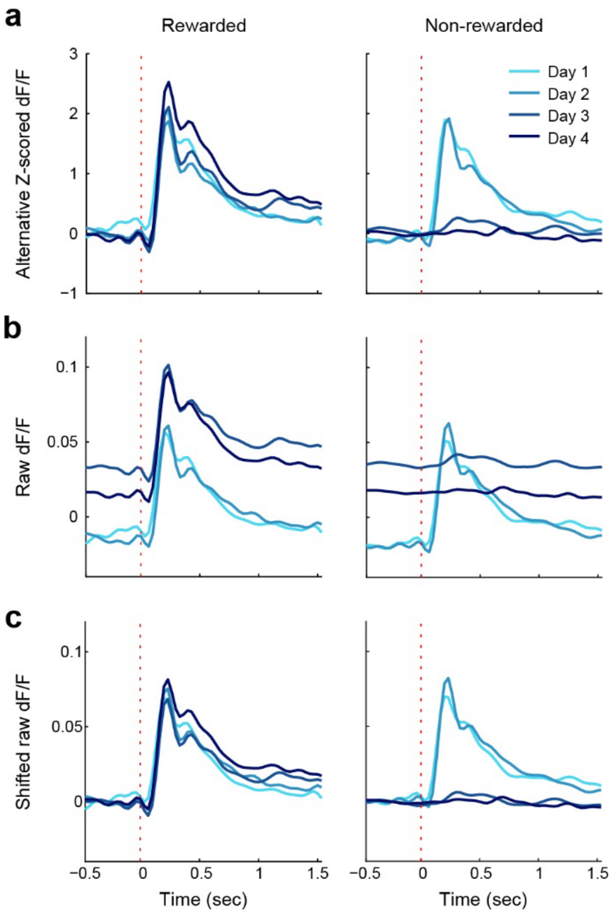
Supplementary Fig. 4. Raw fluorescence for all three sensors. Three sessions from different mice representative of the three sensors used are shown. The header above each plot indicates the animal ID and sensor type. **a** Entire sessions. Active dopamine-modulated fluorescence elicited by 470 nm excitation (blue), and isosbestic fluorescence elicited by 405 nm excitation (red). **b** Same as **a** but with the center portion of the x-axis expanded to show detail over 500 sec. **c** Same as **a** but with the center portion of x-axis expanded to show detail over 20 sec.



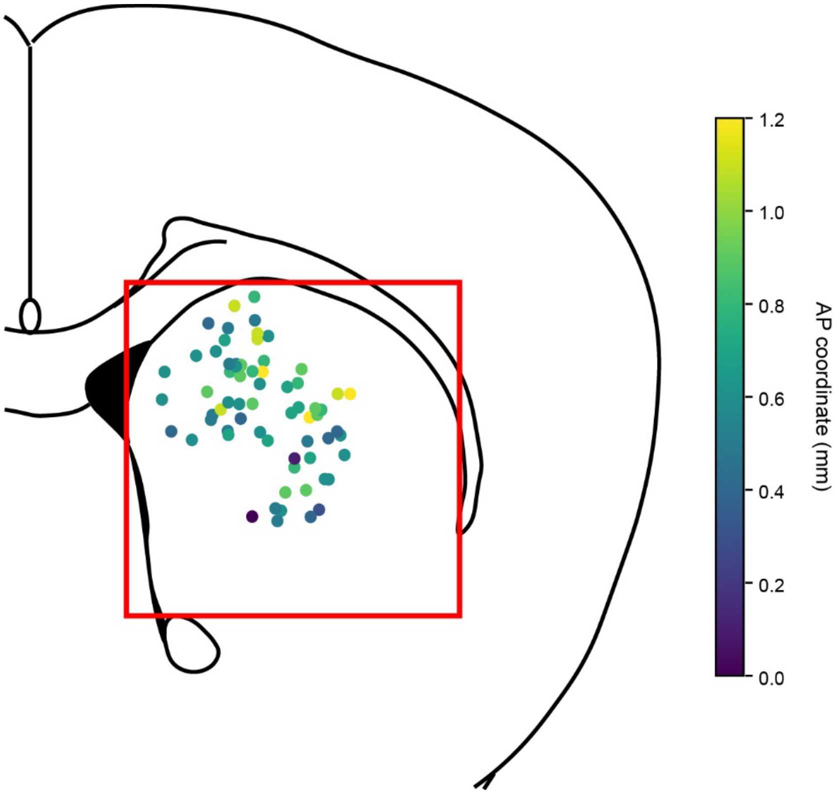
Supplementary Fig. 5. Signal processing pipeline. The same three sessions as **Supplementary Fig. 4** are used here. Each row displays the raw fluorescence signals (left), dF/F (middle) and z-scored dF/F (right) from one session.



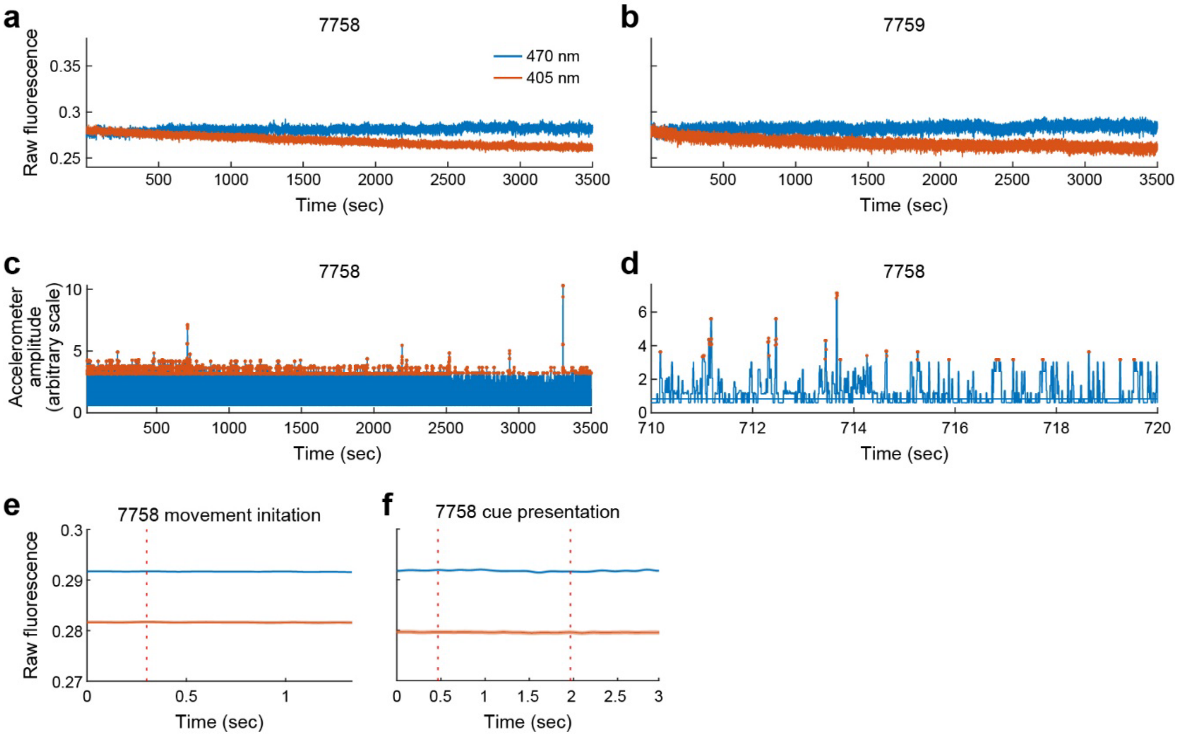
Supplementary Fig. 6. Baseline calibration and choice of reference data for Z-scoring had minimal effects. The same three sessions as in **Supplementary Fig. 4** are shown. **a** In each plot, the green line and shading represent the mean and SEM, respectively, when the standard analysis pipeline was used. The blue line represents the mean when baseline calibration was omitted from the pipeline. **b** Data analysis was the same as in **a**, except that instead of using the entire session to calculate the mean and standard deviation for the z-scoring step, data from the last 1.5 sec before cue onset were aggregated across all trials and used for z-scoring.



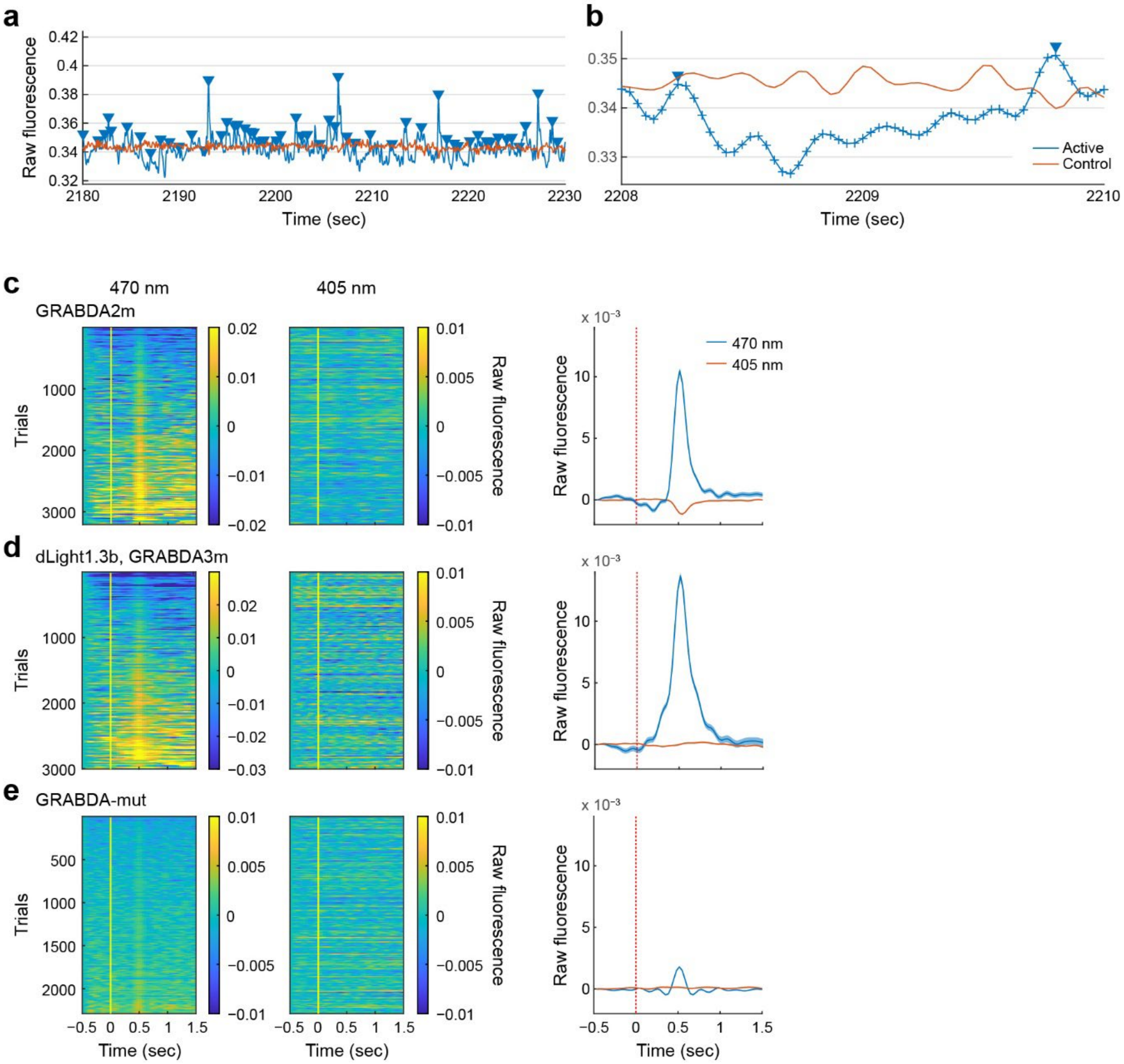
Supplementary Fig. 7. Comparison of normalization methods across sessions. a Same data as in **Fig. 2d** for the Cue discrimination sessions, but re-analyzed using the alternative z-scoring calculation as in **Supplementary Fig. 6b**. **b** Same data as in **a**, but plotted as raw dF/F without any z-scoring. **c** Same data as in **b**, but vertically shifted to align the average of the first 10 points (which are all pre-cue) to 0.



Supplementary Fig. 8. Distribution of recording sites in dorsal striatum. The outline drawing represents the coronal plane at the A-P coordinate of 0.98 mm relative to bregma (adapted from Franklin and Paxinos, 2008). Colored dots indicate the locations of probe tips determined from different histological sections. The color scale represents the A-P position of each dot. The red rectangle outlines the area covered by our standardized coordinate system at the level of the outline drawing, where the left edge is designated "0" and the right edge "1" in the M-L direction, and the top edge is "0" and the bottom edge "1" in the D-V direction. Note that the probe tips located at A-P levels different from the one shown in the outline drawing are projected onto the drawing using our standardized coordinates, not by orthographic projection.



Supplementary Fig. 9. Results from GRAB$_{\text{DA-mut}}$ mice. a, b Raw fluorescence signals from two GRAB$_{\text{DA-mut}}$ mice (ID numbers shown above the plot), recorded with the same methods as those with active GRAB$_{\text{DA}}$ sensors. The active (but in this case disabled) fluorescence trace was excited by 470 nm light (blue) and the control trace was excited by 405 nm light (red). **c, d** Movement peaks detected as in active GRAB$_{\text{DA}}$ sensor mice (see "Additional Control Analyses" for details on peak detection) during the full recording duration (**c**) and during a 10-sec period in the same recording session (**d**). **e** Active (blue) and control (red) raw fluorescence traces aligned to movement initiation and averaged. Shading indicates SEM. The vertical dashed line indicates the movement initiation time. **f** Active(blue) and control (red) raw fluorescence traces aligned to cue onset and averaged. Shading shows SEM. Vertical dashed lines indicate cue onset and cue offset times.



Supplementary Fig. 10. Raw fluorescence in active and control channels aligned to peaks of the active channel. a, b Method for identifying putative dopamine peaks for evaluation of the control signal. Active (blue) and control (red) raw fluorescence traces from one of the active GRAB$_{DA}$ sensor mice during a 50-sec period from a session (**a**) and during a 2-sec period from the same session to make individual samples visible (**b**). Inverted triangles indicate identified peaks. Each sample is marked with a "+" symbol. **c** GRAB$_{DA2m}$ (D2R) sensor data. The left panel displays peak events in the 470 nm channel arranged in ascending order, whereas the middle panel shows the corresponding signals in the 405 nm channel. Note difference in color scale. The right panel depicts raw fluorescence traces averaged across all peaks for the active (blue) and control (red) channels. **d** dLight1.3b and GRAB$_{DA3m}$ (D1R) sensor data, shown as in **c**. **e** Same as **c** and **d**, but for the GRAB$_{DA\text{-}mut}$ control sensor.

Supplementary Table 1. Data on individual mice. "Cue side" designates the side of the initial rewarded cue prior to reversal. The "Cue cond", "Cue disc", and "Rev disc" columns indicate the number of training sessions each mouse received for the cue conditioning, cue discrimination, and reversal discrimination tasks, respectively. "N/A" denotes that the mouse did not reach criterion on the preceding task and did not receive further training.

Mouse ID	Sex	Probe location	Cue side	Cue cond	Cue disc	Rev disc
pa40	M	Medial	R	3	10	16
pa71	F	Medial	L	4	5	23
pp76	M	Medial	L	4	6	0
pp68	M	Lateral	R	6	10	25
pp77	M	Lateral	R	6	9	0
pa74	F	Medial	L	4	13	16
pa03	M	Medial	L	3	3	14
pp69	F	Medial	L	7	10	0
pp82	M	Medial	L	13	3	15
pp84	M	Medial	R	7	7	N/A
pp85	M	Medial	R	7	5	6
pa37	M	Medial	R	7	6	20
pa38	F	Medial	R	3	14	16
pa39	M	Medial	R	5	12	14
pa25	F	Medial	R	3	4	7
pa23	M	Medial	R	3	5	19
pa01	F	Medial	R	3	5	6
pa98	M	Medial	R	5	8	30
pb15	M	Medial	L	9	11	20
pb11	F	Medial	L	3	5	15
pa95	M	Medial	R	10	6	8
pa96	F	Medial	L	3	3	9
pb09	F	Medial	R	3	5	18
pa92	M	Medial	R	3	22	N/A
pa97	F	Medial	R	3	7	12
pb07	M	Medial	L	2	12	8
pa93	F	Medial	R	9	5	14
pb16	F	Lateral	R	4	3	8
pb13	F	Lateral	R	4	5	26
pb12	M	Lateral	R	5	6	16
pa70	M	Lateral	L	12	8	27
pa79	F	Lateral	L	3	5	11
pa72	M	Lateral	L	6	6	19
pa78	F	Lateral	L	3	18	29
pa73	M	Lateral	R	14	27	N/A
pa69	M	Lateral	R	4	3	13
pa66	M	Lateral	L	3	3	7
pa58	F	Lateral	R	3	6	8

pa56	M	Lateral	R	3	8	15
pb04	F	Lateral	L	4	4	15
pb06	M	Lateral	L	3	7	12
pa99	F	Lateral	R	3	14	27
pa91	M	Lateral	L	15	N/A	N/A
pp67	F	Medial	L	5	10	37
pa80	M	Medial	L	11	22	0
pa05	F	Medial	L	5	5	12
pa24	M	Medial	R	3	6	11
pa59	F	Lateral	L	3	8	13
pb31	M	Lateral	L	13	N/A	N/A
pb36	M	Medial	L	12	N/A	N/A
pb32	M	Lateral	L	8	N/A	N/A
pb38	F	Lateral	R	3	7	29
pb37	F	Lateral	L	3	14	30
pb33	M	Lateral	L	3	30	N/A
pb30	M	Lateral	R	12	30	N/A
pb29	F	Lateral	R	3	8	21
pb28	M	Lateral	R	4	8	21
pb35	M	Medial	R	7	24	18
pb39	M	Medial	R	9	30	N/A
pb40	M	Lateral	L	4	24	30
pb42	M	Lateral	L	14	30	N/A
pb43	M	Medial	R	4	4	9
pb44	M	Lateral	R	11	25	N/A
pb46	F	Medial	R	4	22	19
pp38	M	Medial	L	8	16	25
pa67	F	Lateral	R	6	25	27
pa94	M	Medial	L	23	24	N/A

Supplementary Note 1

Trial-averaged Q-learning model

General description

We developed a reinforcement learning model for this task based on the Q-learning algorithm (Ref 2, p. 131), which includes5 states (**Supplementary Fig. 3a**). By arbitrarily assigning a value of 1 to a single reward delivery, we assumed that the cost in effort or energy for making a lick was 0.1, i.e., one reward is equivalent to 10 licks. The state transition diagram in **Supplementary Fig. 3a** pertains to the cue discrimination version of the task. The states are defined strictly in terms of the sensations available to the mouse, specifically whether each cue light is lit or a reward is delivered. We assumed that the mouse can detect whether or not a reward has been delivered, either by smelling it or by hearing the delivery mechanism operating, but this assumption has not been empirically verified. State S5 always transitions back to S1 on the next time step because the only difference between S5 and S1 is whether the non-rewarded cue was lit on the previous time step. S3 transitions back to S1 if the mouse licks because, after the lick, there is no more reward to consume in the model, although it may actually take the biological mouse two or three licks to consume all of the reward. If the mouse waits, however, the model remains in S3 because there is still reward to consume.

In cases where the same action could lead to two different states, the destination state was determined based on the amount of time already spent in the original state, exactly as in the actual experiment. However, since we did not incorporate an internal sense of time passage in the model, we also simplified the model by not randomizing any of the time delays, unlike the actual experiment.

All the task variations shown in **Fig. 1c** can be derived from this state transition diagram by redirecting transition arrows to skip states that are not used in a particular variation and by making reward probabilistic in the probabilistic versions. This approach allows the model to be trained on any desired task order, ensuring that Q values computed for actions in previously experienced states carry over to the next phase of training.

Note that there are no initial or terminal states in this model. We ran the model as a continuous (i.e., non-episodic) task, starting from state S1 at time step 1. Time steps were 0.1 sec in duration, which is the largest time step that can represent a typical mouse's maximum licking rate of approximately 10 Hz.

Comparison of model RPE to experimentally recorded dopamine responses

Model trials were aligned and averaged as done for analysis of experimental trials.

Supplementary Fig. 3b-d illustrates the expected large positive RPE for the initial trials. The model RPE peaks in this task quickly approached their asymptotic value of 0.5, dropping from 0.9 to 0.6 over the first four trials. Therefore, we averaged only the first 3 trials to show the initial response. As expected, after training, the first RPE response shifted one time step (0.1 sec) earlier to cue termination (reward delivery), and the response at the first lick (reward consumption) diminished to about two-thirds of its early trial average. The dopamine responses were clearly quite different.

The model RPE results for cue discrimination demonstrated the classic 'transfer' of response from reward to cue onset (**Supplementary Fig. 3c**). The reward peak initially had twice the RPE value of the cue onset, but by the end of training, the cue onset RPE was almost eight times greater than the reward peak. Such transfer was dramatically absent in the dopamine responses. Model training continued with reversal discrimination (**Supplementary Fig. 3d**), showing a similar pattern: initially, the RPE to the rewarded cue onset was negative, as expected since the model had previously been extensively trained with that same cue as the non-rewarded cue. The response at reward time was positive. After training, the RPE at reward was somewhat smaller, and the RPE at cue onset reached a peak value more than three times that of the reward peak.

Signal processing pipeline and alternative Z-scoring methods

Raw fluorescence signals for all three dopamine sensors

Supplementary Fig. 4 shows the raw traces of our isosbestic control signal (405 nm excitation; red) and putatively dopamine-modulated active signal (470 nm; blue) obtained from a 1-hour session for each of three mice, each prepared with a different dopamine sensor (GRAB$_{DA2m}$, GRAB$_{DA3m}$, and dLight1.3b). Three different time scales are shown to highlight various features. At the whole-session time scale (**Supplementary Fig. 4a**), the peak-to-peak amplitude of the active signal is several times greater than that of the control signal for all three sensors. There is also a small amount of long-term drift, which is typical in photometry recordings. Due to this drift, it is advisable to choose a time point as close to each trial as possible to serve as the reference for zero change in fluorescence. On a 500-sec time base (**Supplementary Fig. 4b**), small fluctuations of a few seconds' duration are apparent in the control signal. At the finest time scale (20 sec, **Supplementary Fig. 4c**), many fluctuations with durations ranging from a few seconds down to about 0.1 sec are observed. Some of these fluctuations appear to correlate positively between the control and active channels, while others show a negative correlation.

These traces demonstrate stable recordings with minimal motion artifact, providing a good signal-to-noise ratio. The slow fluctuations observed in the control trace suggest the possibility of minor motion artifacts or endogenous hemodynamic signaling.

Signal analysis pipeline from raw fluorescence to z-scored dF/F

In **Supplementary Fig. 5**, we show the derivation of our z-scored dF/F signal from the raw fluorescence data. dF was calculated by subtracting the control trace from the active trace (left); the result was then divided by the control trace to yield dF/F (middle). Z-scores were calculated based on the mean and standard deviation of the entire session (right). Single-trial traces of cue response from Pavlovian training day 1, collected from three mice prepared with different sensors, are provided to illustrate each step of the process. In our preparation, we primarily observe rescaling of the ranges.

Effect of baseline calibration

The average effect of performing trial-by-trial baseline calibration was minimal. As shown in **Supplementary Fig. 6a**, for all three sample sessions, the average z-scored dF/F waveform was shifted by less than the width of the SEM, regardless of whether baseline calibration was included in the signal processing.

Reference data selection for Z-scoring

There are several different methods for calculating z-scores used in the published literature. They vary based on the data selected to compute the mean and standard deviation used in the z-score formula. In the present study, we chose to use a method based on the entire data recording for the session, including both inter-trial interval and task activity. A potential drawback of this approach is that significant changes in response size during the task could be normalized out by the z-scoring. To address this, we repeated some analyses using an alternative data selection method to calculate the mean and standard deviation for the z-score. Specifically, we used all the samples from the last 1.5 sec of the inter-trial interval (i.e., the baseline period ending at cue onset) preceding each trial. These samples were aggregated into one set, and the mean and standard deviation of this aggregated set were computed. **Supplementary Fig. 6b** shows the same results as **Supplementary Fig. 6a** but computed with z-scores based on this alternative data selection rather than the entire sessions. The two sets of plots are difficult to distinguish by eye, except that the mouse with the dLight1.3b sensor shows a visibly taller peak using the alternative z-score (3.5 standard deviations) compared to the whole session-based z-score (3.3. standard deviations). Thus, the choice of data for calculating z-scores had a remarkably small effect on the final result.

Finally, we recalculated the data shown in **Fig. 2d** using the alternative z-score method

(**Supplementary Fig. 7a**) and without any z-scoring (**Supplementary Fig. 7b**). The vertical shifts between sessions in **Supplementary Fig. 7b** make it difficult to compare this figure directly to **Fig. 2d**. To address this, we re-plotted the data from **Supplementary Fig. 7b** after subtracting the average of the baseline period, and these results are shown in **Supplementary Fig. 7c**. Comparing **Supplementary Fig. 7a** and **b** to **Fig. 2d**, we still observe that at the start of cue discrimination training, both cues produced a slowly decaying dopamine response that extended well beyond the initial peak. As training proceeded, the response to the reward-predicting cue became larger and more sustained, while the response to the non-reward-predicting cue was nearly abolished by the third session.

Controlling for motion artifacts

Lack of high amplitude transients in $\text{GRAB}_{\text{DA-mut}}$ mice

Two mice were injected with $\text{GRAB}_{\text{DA-mut}}$ (140555-AAV9, Addgene) instead of the functional GRAB_{DA}, and we added an accelerometer to our recording chamber to detect movements. The sensitivity of the accelerometer to the mice's movements was assessed by running another mouse on the standard training tasks with simultaneous video recording. A number of movement incidents were identified by examining the recorded video, and it was confirmed that each visible movement was consistently accompanied by an accelerometer signal. Most movements were tail movements, with occasional face or forepaw movements.

The two $\text{GRAB}_{\text{DA-mut}}$ mice (7758 and 7759) were then implanted in centromedial and centrolateral striatum, respectively, and photometry data were collected with simultaneous accelerometer recordings. The raw traces using the $\text{GRAB}_{\text{DA-mut}}$ sensor replicated the stability and other general features observed with the active dopamine sensors (see "Additional Control Analyses" section for corresponding analyses of the active dopamine sensor data), with one exception: as expected, the $\text{GRAB}_{\text{DA-mut}}$ fluorescence evoked by 470 nm excitation (blue line) did not show substantially larger signals than those evoked by the 405 nm (red line) nominally isosbestic excitation (**Supplementary Fig. 9a, b**).

The absence of high-amplitude transients at 470 nm (blue) is presumably due to the fact that the $\text{GRAB}_{\text{DA-mut}}$ sensor is not modulated by dopamine transients. Although the slow deflections in the blue trace appear slightly larger than those in the red trace, they are still smaller than the high-frequency noise deflections present in both traces. We therefore conclude that the control signal used in our study is free from any issues and effectively serves its intended purpose.

Lack of transients corresponding to movements in $\text{GRAB}_{\text{DA-mut}}$ mice

Further evidence of the quality of the original control signal comes from traces averaged around detected movements. In **Supplementary Fig. 9c and d**, we present a trace of the vector magnitude of the 3-dimensional accelerometer signal for an entire session, with significant movement peaks marked. The accelerometer data were collected concurrently with the mouse's performance in a Pavlovian session using a head-fixed apparatus. The x, y and z values of the accelerometer were combined into an accelerometer amplitude, calculated as the square root of the sum of the squares. Movement detection was defined as instances where the amplitude exceeded the 90th percentile threshold. Detected movements were then grouped if they occurred within 333 msec of each other, with the first instance considered the start of the movement. Traces were extracted from both mutant sensor and control signals and aligned to the start time of each movement episode. The traces were then averaged to reduce noise and are shown in **Supplementary Fig. 9e**. These traces display the average of the raw fluorescence signals without any calibration or normalization.

A similar analysis was performed by aligning the traces on each cue presentation. The results are shown in **Supplementary Fig. 9f**. Once again, no significant movement artifacts or cue-evoked signals were observed in either channel. This further confirms that the control channels effectively served their purpose in calibration for our study.

Evaluation of nominally isosbestic channel as a control signal in main data set: data preparation

Initially, peaks with a minimum width of 5 data points were identified in the raw 470 nm channel (inverted triangles in **Supplementary Fig. 10a, b**). A 2-sec window, spanning 1 sec before and after each peak, was then extracted, totaling 60 samples per peak. The value of the first sample in the window was subtracted from all samples in the window, ensuring that each window began with a sample value of zero. A similar calibration process was applied to the 405 nm channel, based on the peak positions detected in the 470 nm channel. The data were then sorted according to the peak amplitudes in the 470 nm channel to ensure the inclusion of all activated instances, which occur continuously and spontaneously rather than being event-specific. Finally, the relationship between the signals in the two channels was examined.

Supplementary Note 2

The GRAB$_{DA2m}$ (D2R) sensor data consistently exhibited the pattern shown in **Supplementary Fig. 10c**. The consistently timed negative deflection and amplitude dependency observed at 405 nm indicates a small but systematic negative covariation between the 405 nm isosbestic control measurements and the 470 nm dopamine-modulated measurements in the $\text{GRAB}_{\text{DA2m}}$ sensor. This covariation may suggest that the true isosbestic point in our material is slightly closer to 470 nm than the wavelength we used for our control channel. However, since the contamination of the 405 nm control signal is small (~0.001) relative to both the size of the deflection in 470 nm trace (~ 0.01) and the absolute size of the 405 nm signal (~0.3), it should not significantly affect the interpretation of the results. When calculating dF/F, the contaminating deflection in the control channel will add roughly 10% to the amplitude of the peak dF. Dividing by F will add another approximately 0.3%, but this is negligible compared to roughly 10%. A precise quantitative analysis would require compensation for this ~10% systematic error, but we do not attempt such precision in the current paper.

The dLight1.3b and $\text{GRAB}_{\text{DA3m}}$ (D1R) sensors all demonstrated the pattern shown in **Supplementary Fig. 10d**. In these cases, the contamination introduced into the control signal is smaller than the confidence limits on the 470 nm signal, which will be subtracted from it to compute dF. Therefore, this contamination is negligible. This suggests that 405 nm is likely closer to the true isosbestic point for these sensors than it is for $\text{GRAB}_{\text{DA2m}}$. In any case, only the $\text{GRAB}_{\text{DA2m}}$ recordings appear to be inflated by as much as 10%.

For completeness, we performed the same analysis on the $\text{GRAB}_{\text{DA-mut}}$ mice, which also exhibited the pattern shown in **Supplementary Fig. 10e**. The data from these mice were prepared using the same method as described earlier. This finding confirms that the control channel (405 nm excitation) showed no contamination from the active channel (470 nm excitation). The small signal peaks detected in the active channel likely reflect the influence of other time-varying factors, aside from dopamine, on this sensor.