Reservoir of decision strategies in the mouse brain

Fanny Cazettes[1*], Masayoshi Murakami[1,2],
Alfonso Renart[1,†*], Zachary F. Mainen[1,†*]

[1]Champalimaud Centre for the Unknown, Lisbon, Portugal.
[2]Department of Neurophysiology, University of Yamanashi, Japan.

*__Correspondence to__:
fanny.cazettes@neuro.fchampalimaud.org,
zmainen@neuro.fchampalimaud.org,
alfonso.renart@neuro.fchampalimaud.org.
†Equal contribution.

Abstract

Decision making strategies guided by observable stimuli and those that also require inferences about unobserved states have been linked to distinct computational requirements and neural substrates. Here, we formulate a model based on temporal integration and reset that incorporates both strategies into a unified family of decision algorithms. We show, using recordings from the frontal cortex of mice performing a foraging task, that the entire family of algorithms can be simultaneously decoded from the same neural ensemble, regardless of the one concurrently executed by the mice. Thus, using multiplexed integration, the cortex may avoid premature commitment to a single algorithm and maintain multiple decision strategies in parallel.

Introduction

An adaptive strategy to control behavior is to take actions that lead to good outcomes given that the environment is in a particular state. It is believed that different circuits and computations are involved in this process depending on whether the relevant environmental state is observable and unambiguous — leading to “stimulus-bound” strategies — or whether it is latent and requires making inferences using knowledge of the causal structure of the world (*1*, *2*). Stimulus-bound agents learn by gradually using outcome information across trials to shape the value of actions, whereas, when inference is needed, frontal structures are believed to generate representations of latent states which enable faster and more flexible behavioral adaptation (*3*, *4*). From a purely computational point of view, however, it is conceivable that these two types of strategies could be unified based on a common set of primitives. In fact, both types of strategy critically rely on temporal accumulation of evidence about the relevant action outcomes (*2*, *5*, *6*). At a mechanistic level, recurrent neural networks have been shown to provide accurate descriptions of decision-making in frontal and motor cortical networks (*7*, *8*), including the ability to compute representations that allow simultaneous decoding of multiple relevant readouts (*9*–*11*). We studied this problem using a recently-developed foraging task for mice which admits different solution strategies (*2*, *12*). In one strategy, decisions to leave a foraging site are based on a variable that embodies an inference about a hidden state indicating that resources are still available. Alternatively, decisions could simply be based on how much reward mice have recently experienced at the site. Here, we show that a generative model can be defined for a reservoir of decision variables of which the two just highlighted are particular examples, and provide evidence that the whole reservoir can simultaneously be decoded from neural ensembles in the mouse frontal cortex.

Probabilistic foraging task

In our task, a head-fixed mouse collected rewards at a virtual foraging site by licking from a spout (Fig. 1A; Fig. S1). During a foraging bout, either a fixed amount of reward or nothing was delivered for each detected lick. At any time, the mouse could choose to continue licking or, if faced with a depleted foraging site, give up and explore a new site by starting to run. By design, reward delivery followed a schedule that admitted a particular optimal strategy. There were two virtual foraging sites only one of which was active at a given time and would deliver reward with a probability of 0.9 after each lick. The identity of the active site had also a probability of 0.3 of switching after each lick. Therefore, the optimal strategy was to infer the latent state corresponding to which port was currently active. This could be done by temporally accumulating consecutive failures with a complete reset upon receiving a reward (Fig. 1B), and leaving the current site when this decision variable reached a given threshold (*2*). This is because a failure to receive reward corresponded to either an unlucky attempt in the active state or to evidence in favor of the hypothesis that the active state had switched, whereas a single reward always signalled the active state with certainty. Yet, in principle, mice could seek reward by using any number of strategies based on the sequence of rewarded and unrewarded licks, such as staying longer at the site when reward rate is high (*2*), which would be consistent with a stimulus-bound algorithm that uses the net negative ‘value’ of the site (the difference between failures and reward) as the relevant decision variable (Fig. 1C).

After several days of interaction with this setup (n = 13 ± 5 days; mean ± s.d.; Fig. S2), mice (n = 21) learned to exploit each site for several seconds (Fig. 1D,E). Considering the last sessions of the training, we examined how well the different decision variables supporting the inference-based and the stimulus-bound strategies predicted in each session the mouse’s choice to switch sites. Specifically, we used regularized logistic regression to model the probability that each lick (n = 2,882 ± 1,631 licks per session; mean ± s.d.) was the last one in the bout using the stimulus-bound vs. inference-based decision variables as predictors (Fig.1F, Methods). We found that decision variables were used to different extents in different sessions, with mice largely relying on the optimal strategy in some, while in others relying on the accumulated value of the foraging site (Fig. 1G).

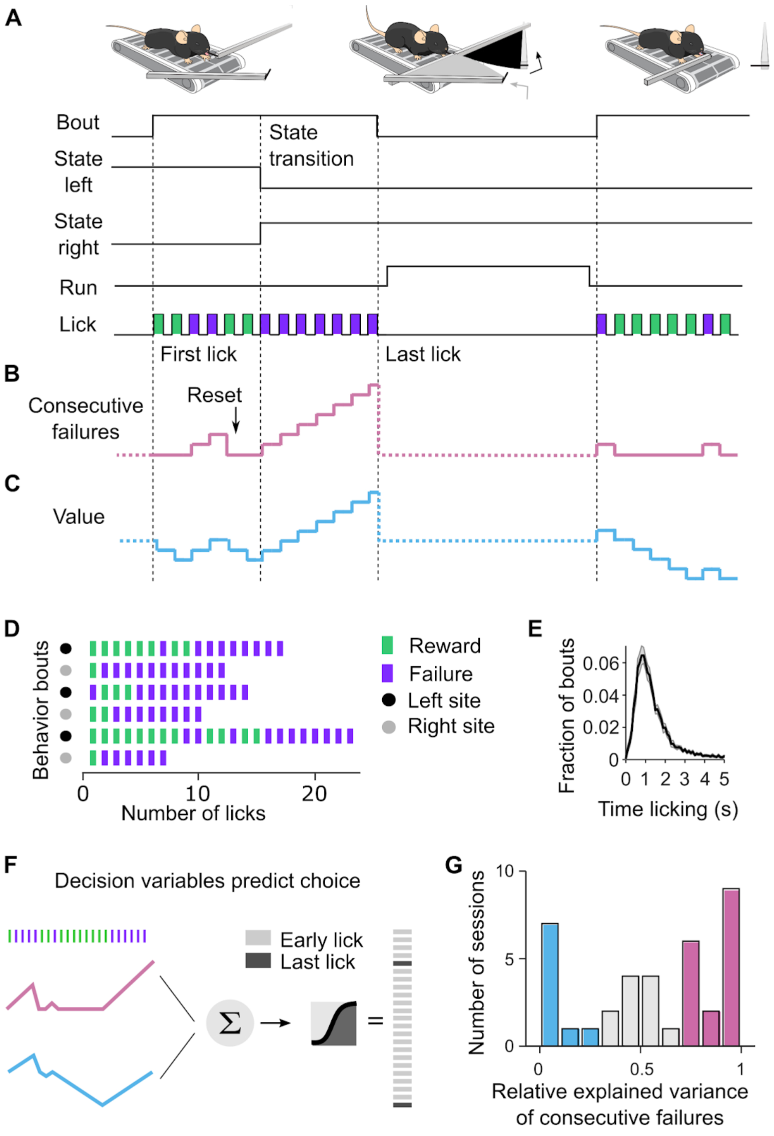

Fig. 1. Probabilistic foraging task. (A) A head-fixed mouse placed on a treadmill chooses to exploit one of the two foraging sites (two movable arms on each side of the treadmill). A bout of behavior consists of a series of rewarded and unrewarded licks (action outcomes) at one of the sites. When a site is in active state, the probability of each lick being rewarded is 90% and each lick is associated with a 30% probability of state transition. Independently from state transition, animals can choose to switch between sites at any time by running a set distance on the treadmill. During site-switching, the spout in front moves away and a distal one moves into place. (B) The decision variable that the mouse needs to compute to optimally solve the task. (C) Alternative decision variable supporting a stimulus-bound strategy: the negative 'value'. (D) Example sequences of observable events during different behavior bouts. (E) Histogram of bout duration (mean ± s.e.m across mice; n = 21). (F) Illustration of the logistic regression method for predicting the choice of the mouse from the two different decision variables. (G) Distribution of the relative explained variance from the logistic regression of the decision variable 'consecutive failures' relative to the decision variable 'value' across sessions (n = 37). Blue are sessions where the decision variable 'value' explain most of the variance of the choice behavior, while pink corresponds to sessions where the decision variable 'consecutive failures' explain most of the variance of the choice behavior.

Unified family of decision strategy

The decision variables underlying the inference-based and stimulus-bound strategies both rely on the accumulation of failures across time, but they differ in how they treat rewards, which are accumulated with the opposite sign to failures in one case, and reset the count of failures in the other. The different effects of reward on the two decision variables can be conceptualized as an adaptive, outcome-dependent feedback gain on the running count. For instance, if we refer to the running count after the *t*-th lick as x_t and to the outcome of the next lick as ot*+1* (equal to plus or minus 1 if the outcome is a reward or a failure respectively), then we can write the update rule compactly as

$$x_{t+1} = g(o_{t+1})\, x_t \;+\; o_{t+1}$$

with $g(o_{t+1} = 1) = 0$ and $g(o_{t+1} = -1) = 1$ for the inference-based strategy and $g(o_{t+1} = 1) = g(o_{t+1} = -1) = 1$ for the stimulus-bound decision variable. This realization suggests that a common generative model can generate these two different decision variables by adjusting certain model parameters. The simplest such model contains two discrete outcome-dependent parameters: one is a gain factor which specifies whether the running count should be reset or accumulated by each outcome – a non-linear operation – and the other specifies how each outcome linearly contributes to the resulting running count, which in general could be positive, negative, or zero (leaving it unaffected; Fig. 2A, Methods). Each specification of these two parameters leads to a different decision variable. This model thus describes, within a single algorithmic framework, the computations necessary to generate, not only the two decision variables considered so far, but also other decision variables relevant for a variety of commonly studied behavioral tasks, such as for instance a 'global count' (accumulated number of outcomes) decision variable, or a cumulative rewards decision variable (Fig. 2B, C). Although not essential in the present task, these particular 'stimulus-bound' strategies could be useful in a different behavioral context (*5*, *13*).

As it is, the model generates a space of time series of rank 8 (see Methods). However, the effective dimensionality of the space of decision variables depends on the temporal statistics of the outcome sequences (which are a function of the reward and state-transition probabilities). Running principal component analysis on the space of decision variables generated from the outcome sequences experienced by the mice, we found that the effective dimensionality was approximately 4 under our task conditions (Fig. 2D, Methods).

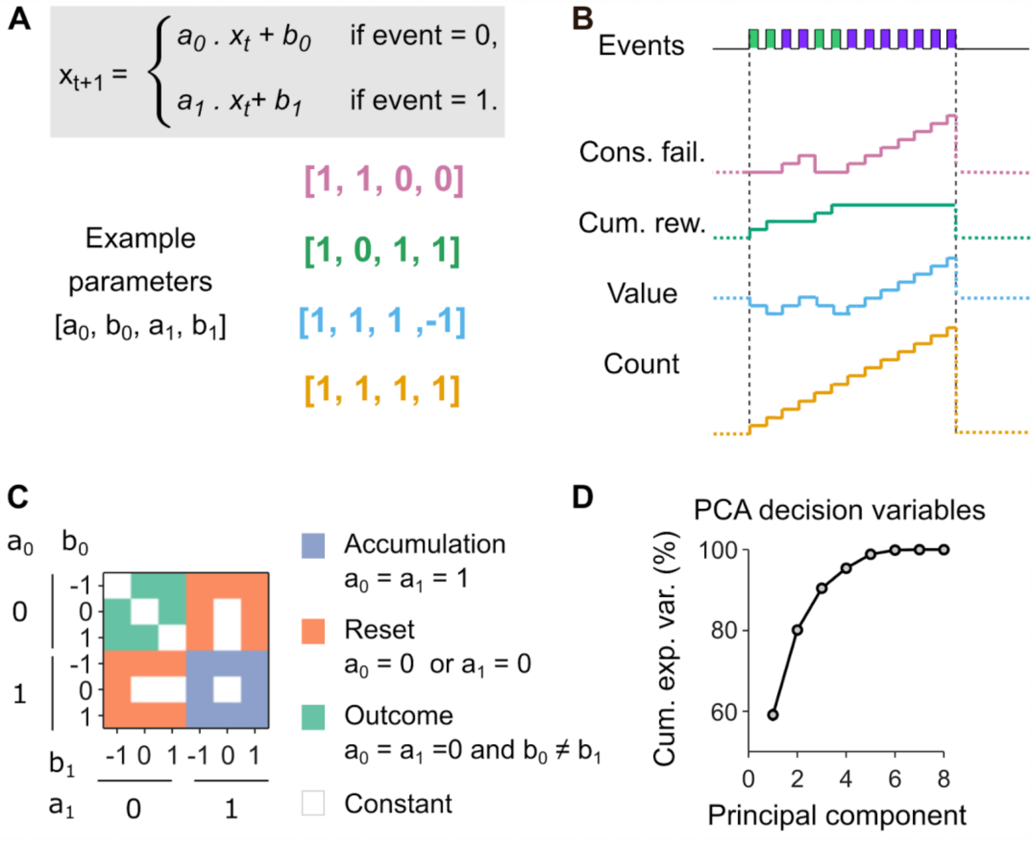

Fig. 2. A unified family of decision strategy. (A) A model with four parameters generates different time series by accumulating, resetting or ignoring each possible event (reward or failure). Example set of parameters yielding the example decision variables in (B). (C) The parametric model yields a combination of 36 time series. (D) We ran principal component analysis on the space of time series in (C) generated from the outcome sequences experienced by the mice and reported the cumulative sum of the percentage of the total variance explained by each principal component.

Neural representation of decision strategies

To examine the neural basis of selective evidence accumulation, we used Neuropixels electrode arrays (*14*) , which are single shank probes with hundreds of recording sites that allowed registering the activity of large ensembles of neurons (n = 145 ± 32 per session; mean ± s.d.) in two regions of the frontal cortex, the secondary motor cortex and the orbitofrontal cortex, while mice performed the task (Fig. 3A,B; Fig. S3). We considered the instantaneous response patterns of isolated neurons in small time windows around each lick (Fig. 3C, Methods). Whereas we observed heterogeneous task-related activity in many single neurons (Fig. S4), we focused our analysis on representations at the population level from each single recording session (one recording per mouse; n = 7). We used cross-validated regression-based generalized linear models (GLM; see Methods) to decode the instantaneous magnitude of different decision variables – all during each bout (n = 223 ± 119 bouts per session; mean ± s.d.) and on a on a lick-by-lick basis – from the responses of neurons in the frontal cortex (Fig. 3D).

The data from Fig. 3A-G are from a single recording session during which the mouse relied mainly on the inference-based strategy (relative variance explained for consecutive failures = 81%, calculated as in Fig. 1G). Here, 140 neurons were simultaneously recorded from the frontal cortex. We found that frontal populations reflected temporal accumulation of failures and reset of integration upon observation of different events (i.e. after a reward and at the beginning of the next bout; Fig. 3E, top & Fig. S5), which are the computations necessary to support the inference-based strategy. Surprisingly, we also found that we could simultaneously decode various alternative decision variables that relied on different computations (Fig. 3E). One possible explanation for this finding is that the mouse actually used all the decision variables to a similar extent to drive its behavior. To investigate whether this was the case, we again tested how well each neural projection predicted the animal's choice to switch sites by using logistic regression to model the probability that each lick (1853 licks in this example session; overall n = 2,533 ± 1,524 licks per session; mean ± s.d.) was the last one in the bout (Fig. 3F). We found that some neural projections were highly weighted, such as the ones associated with the consecutive failure and the value, while the remaining projections had null weights. Remarkably, however, there was overall no evidence that, across recording sessions, the accuracy of the representation of a particular decision variable was correlated with its predictive power for the behavior (Fig. 3G, $r^2 = 0.03$, $p = 0.39$).

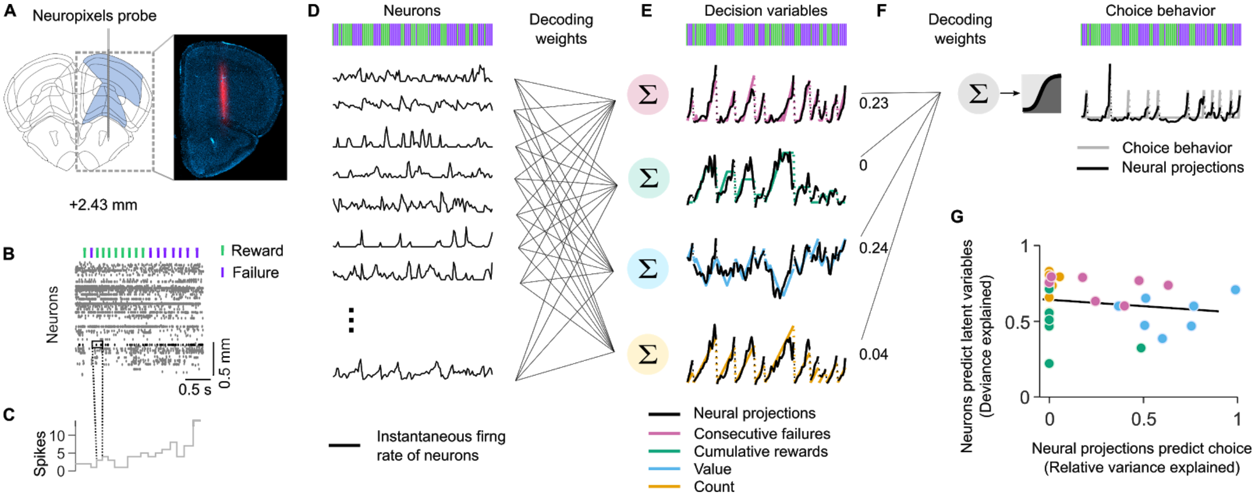

Fig. 3. Neural representation of decision strategies. (A) Schematic target location of probe insertion and example histology of electrode track (see Methods). (B) Example raster plot of 140 simultaneously recorded neurons from the frontal cortex. Lick-outcome times are indicated by the green (reward) and purple (failure) dashes. (C) Binned response profile of an example neuron. For all analyses, otherwise noted, we averaged for each neuron the number of spikes into bins by considering a 200 ms window centered around each lick (see methods). (D) The regression models take as predictors the activity of simultaneously recorded neurons (black traces) and derive a set of decoding weights to predict the decision variable. (E) Predictions of the models (black traces are the weighted sums of neural activity) overlaid onto the decision variables (color traces). (F) Illustration of the logistic regression method for predicting the choice of the mouse from the neural projections (black traces in E) of the different decision variables (color traces in E). The decoding weights associated with the different decision variables are shown for this particular example. (G) Correlation between the neural representations of different decision variables (color coded) and their roles for behavior (indicated by the relative variance explained in abscissa). Each dot corresponds to a particular decision variable from a given recording session. The linear regression is reported in black.

A basis set for decision variables

The above results indicate that the mouse's brain does not only compute a single decision variable tailored to the current behavioral strategy but also computes in parallel alternative decision variables, including both what would be considered inference-based and stimulus-bound strategies. A possible concern with this interpretation is that the alternative decision variables we considered might be represented only by virtue of being similar to the ones reflected behaviorally. Although the computations underlying the multiple decision variables are different, for the particular sequences of rewards and failures experienced by the mice, the decision variables themselves can sometimes be highly correlated (Fig. 4A).

To address this issue, we first note that all non-trivial time series produced by the generative model can be expressed as linear combinations of four sequences, which we refer to as basis elements (BEs, Fig. 4B; Methods). The two BEs involving reset describe integration of failures and reset by rewards (the consecutive failures) and vice-versa (Fig. 4B, top). The two BEs for accumulation without reset were upwards integration of both rewards and failures (equivalent to ‘count’) and integration upwards of rewards and downwards of failures (equivalent to ‘value’; Fig. 4B bottom). In all cases, the BEs are reset at the beginning of a bout.

Next, we proceeded to investigate the extent to which each of the four BEs is well represented independently of the others. To do so, we focused on a particular subset of outcome sequences (unlike Fig. 3, where we considered all outcome sequences experienced by the mice). Namely, we considered outcome sequences starting at the beginning of a bout, and consisting of a series of *N* consecutive rewards followed by a sequence of *N-1* failures, for several values of *N*. For these outcome sequences the two BEs involving reset are clearly distinct, and the two BEs for pure accumulation are orthogonal (Fig. 4B), so an accurate representation of one particular BE cannot by itself account for a representation of the other BEs.

The time series associated with each BE could be accurately decoded from the neural activity in the frontal cortex (Fig. 4C,D) with better accuracy for time series with higher variance (Fig. 4D). Moreover, linear combinations of the two BEs of accumulation of rewards and failures (with equal and different signs) produce time series consisting of ramps and persistence, which could also be decoded (Fig. 4E,F). Yet, for these rewards-followed-by-failures outcome sequences, the BE consisting of persistence for rewards and accumulation of failures and the consecutive failures were confounded. Nevertheless, focusing on specific subsets of the existing set of outcome sequences it was possible also to establish that these two BEs were indeed independently represented (Fig. S6).

Because the individual BEs are well represented, the frontal populations should also be able to represent each decision variable that the generative model can produce when fed with the sequences of action outcomes experienced by the mouse. As predicted, this repertoire could indeed be decoded from the population activity (Fig. S7).

All the decoding analyses were cross-validated to ensure that the data were not overfit, but as a further check that these results reflect true properties of frontal cortex, we compared the decoding quality from an area that we reasoned should not be involved in state inference in this task, the olfactory cortex (Fig. 4G). We found that every single BE (both task and non-task related) could be better decoded from neurons in the frontal cortex than from neurons recorded simultaneously in the olfactory cortex (Fig. 4H). As a final control, we also tested whether the movement of the mice would predict the decision variables. However, we found no strong links between the decision variables and facial and forelimb movements that may be associated with frontal cortical activity (Fig. S8). Together with previous studies (*1*, *2*, *15*, *16*), these results point to the frontal cortex as a privileged location for the representation of decision variables.

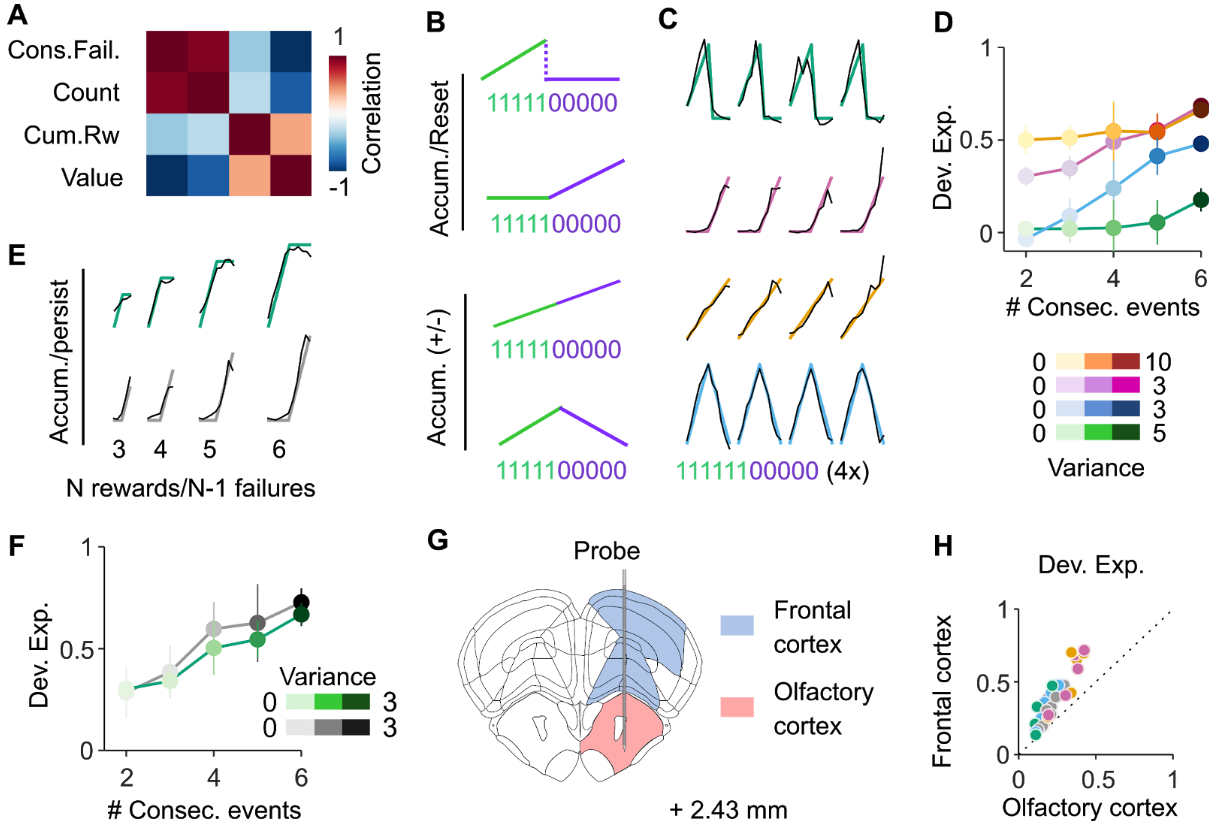

Fig. 4. A basis set for decision variables. (A) Correlation matrix of the four decision variables in Fig. 3. (B) Four BEs involving reset and accumulations of rewards and failures. (C) Four example bouts (columns) of population activity (black traces) projected onto the dimensions that best predict the trajectory of the four BEs (color traces). Only sub-sequences of consecutive rewards followed by consecutive failures were selected in order to orthogonalize the time series (~5% of bouts with N = 6). (D) Decoding quality of each BE and the respective variance of the different time series (median ± m.a.d.; n = 7; light to dark colors for small to large variances). (E) Examples of population activity (black traces) projected onto the dimensions that best predict the trajectory of BEs that accumulate rewards and persist with failures (top) and vice-versa (bottom). (F) The decoding quality of the 'accumulate and persist' BEs and the respective variance of the different time series (median ± m.a.d.; n = 7; light to dark colors for small to large variances). (G) Simultaneous recording in the frontal cortex and in a control region: the olfactory areas. (H) The BEs were better decoded from the frontal cortex than from olfactory areas (number of neurons randomly matched between the two regions; n = 49 ± 25 neurons per session, mean ± s.d.). Each color dot represents a given BE for a given recording, grey dots represent the observable events (i.e., decoding of reward and failure).

Discussion

Here, we explored the capacity of the frontal cortex to implement different kinds of decision strategies, that is, to create different algorithms for generating a diversity of decision variables. We studied this in the context of a foraging task whose solution required mice to process streams of successful and unsuccessful foraging attempts (licks) executed over several seconds. We found that mice deployed not one but multiple processing strategies, all of which could be read out from populations of neurons in the secondary motor cortex and orbitofrontal cortex. Moreover, we found that these different strategies were actually implemented simultaneously and in parallel within the same neural populations and that the neural availability of alternative strategies was nearly completely independent of the actual currently-deployed decision strategy. These results thus reveal a hidden computational reserve in the frontal cortex, which is not revealed in behavior and can only be uncovered through neural recordings.

The different decision variables in this hidden computational reserve are “mixed” but can be recovered through linear decoding. Although multiplexed neural codes have been observed previously (*7*, *17–19*), our results establish that the kind of information that is multiplexed is not limited to representations of observable events, but also includes temporally extended computations spanning several seconds.

One such computation is accumulation of evidence which, through its intimate relationship with posterior beliefs (*20*, *21*), constitutes an essential computation for statistical inference, and has therefore been implicated in a variety of decision-making and reasoning tasks (*6*, *22–26*). Accumulation (possibly temporally discounted) of action outcomes also underlies several reinforcement learning algorithms (*4*, *27–29*). Although less attention has been devoted to reset-like computations (but see (*30*)), they are also essential for inference when certain observations specify a state unambiguously (*2*). Our results show that these computations can be cast into a single parametric generative model capable of producing a set of decision variables embodying a variety of stimulus-bound and inference-based strategies. Furthermore, the fact that there appears to be no correlation between the accuracy with which a particular decision variable is represented and the extent to which its associated strategy explains mice behavior, suggests that this generative model is run by default for different parameter settings in the frontal cortex and associated areas.

It is very likely that the set of decision variables we identified reflects a lower bound on the richness of the actual model family instantiated by the cortex. For instance, we considered a subset of cases in which all or none integration occurs, but the model is readily extensible, through analog parameter values, to produce leaky integration with different time constants. Functionally, the ability of frontal populations to simultaneously implement different algorithms supporting several useful decision strategies would allow adapting behavior to new contingencies by simply modifying linear readouts of these frontal neural populations based on feedback (*31*, *32*) and without the need to implement new computations, consistent with the framework of ‘reservoir’ computing (*9*, *10*, *33*, *34*).

Our finding also speaks to the debate on the nature of serial processing limitations in the brain. While it has been shown that limitations apply in some kinds of evidence accumulation tasks

(*35–37*), here we show in a different, but ecologically important, setting that some forms of evidence accumulation can proceed in parallel. Thus, our findings are consistent with proposals favoring parallel integration (*38–40*) and with models that place serial constraints on behavior close to the specification of the timing of action (*38*, *41*).

Materials and Methods

Experimental methods

Animal subjects. A total of 21 adult male and female C57BL/6J mice (2-9 months old) were used in this study. All experimental procedures were approved and performed in accordance with the Champalimaud Centre for the Unknown Ethics Committee guidelines and by the Portuguese Veterinary General Board (Direco-Geral de Veterinria, approval 0421/000/000/2016). During training and recording, mice were water-restricted (starting 5 to 10 days after head-bar implantation), and sucrose water (10%) was available to them only during the task. Mice were given 1 mL of water or 1 gram of hydrogel (Clear H2O) on days when no training or recording occured or if they did not receive enough water during the task.

Surgery and head-fixation. All surgeries used standard aseptic procedures. Mice were deeply anesthetized with 4% isoflurane (by volume in O2) and mounted in a stereotaxic apparatus (Kopf Instruments). Mice were kept on a heating pad and their eyes were covered with eye ointment (Vitaminoftalmina A). During the surgery, the anesthesia levels were adjusted between 1% and 2% to achieve 1/second breathing rate. The scalp was shaved and disinfected with 70% ethanol and Betadine. Carprofen (non-steroidal anti-inflammatory and analgesic drug, 5 mg/kg) was injected subcutaneously. A flap of skin (less than 1cm^2) was removed from the dorsal skull with a single cut and the skull was cleaned and dried with sterile cotton swabs. The bone was scraped with a delicate bone scraper tool and covered with a thin layer of cement (C&B super-bond). Four small craniotomies were drilled (HM1 005 Meisinger tungsten) between Bregma and Lamba (around −0.5 and −1 AP; ±1 ML) and four small screws (Antrin miniature specialities, 000-120x1/16) previously soaked in 90% ethanol, were inserted in the craniotomies in order to stabilize the implant. The head-bar (stainless steel, 19.1 × 3.2 mm), previously soaked in 90% ethanol, was positioned directly on top of the screws. Dental cement (Tab 2000 Kerr) was added to fix the head-bar in position and to form a well around the frontal bone (from the head-bar to the coronal suture). Finally an external ground for electrophysiological recording (a male pin whose one extremity touched the skull) was cemented onto the head-bar.

Behavioral apparatus. Head-fixed mice were placed on a linear treadmill with a 3D printed plastic base and a conveyor belt made of Lego small tread links. The running speed on the treadmill was monitored with a microcontroller (Arduino Mega 2560), which acquired the trace of an analog rotary encoder (MAE3 Absolute Magnetic Kit Encoder) embedded in the treadmill. The treadmill could activate two movable arms via a coupling with two motors (Digital Servo motor Hitec HS-5625-MG). A lick-port, made of a cut and polished 18G needle, was glued at the extremity of each arm. Water flowed to the lick-port by gravity through water tubing and was controlled by calibrated solenoid valves (Lee Company). Licks were detected in real time with a camera (Sony PlayStation 3 Eye Camera or FLIR Chameleon-USB3) located on the side of the treadmill. Using BONSAI (*42*), an open-source visual programming language, a small squared region of interest was defined around the tongue. To detect the licks a threshold was applied to the signal within the region of interest. The behavioral apparatus was controlled by microcontrollers (Arduino Mega 2560) and scientific boards (Champalimaud Hardware platform), which were responsible for recording the time of the licks and the running speed on the treadmill, and for controlling water-reward delivery and reward-depletion according to the statistics of the task.

Task design. In the foraging task, two reward sites, materialized by two movable arms, could be exploited. Mice licked at a given site to obtain liquid reward and decided when to leave the current site to explore the other one. Each site could be in one of two states: “ACTIVE”, i.e. delivering probabilistic reward, or “INACTIVE”, i.e. not delivering any reward. If one of the sites was “ACTIVE”, the other one was automatically “INACTIVE”. Each lick at the site in the “ACTIVE” state yielded reward with a probability of 90% , and could cause the state to transition to “INACTIVE” with a probability of 30%. Licks could trigger the state of the exploited site to transition from “ACTIVE” to “INACTIVE”, but never the other way around. Importantly, this transition was hidden to the animal. Therefore, mice had to infer the hidden state of the exploited site from the history of rewarded and unrewarded licks (i.e., reward and failures). We defined ‘behavioral bout’ the sequence of consecutive licks at one spout. A tone (150 ms, 10 kHz) was played when one of the arms moved into place (i.e., in front of the mouse) to signal that a bout could start. At the tone, the closed-loop between the motors and the treadmill decoupled during 1.5 s or until the first valid lick was detected. During this time mice had to “STOP”, i.e. decrease their running speed for more than 250 ms below a threshold for movement (3 cm/s). Licks were considered invalid if they happened before “STOP” or at any moment after “STOP” if the speed was above the threshold. If a mouse failed to “STOP”, “LEAVE” was triggered by reactivating the closed-loop after 1.5 s, which activated the movement of the arms (the one in front moved away and the other moved into place). Mice typically took around 200 ms to “STOP” and initiate valid licking. During the licking periods, each lick was rewarded in a probabilistic fashion by a small drop of water (1 μl). The small reward size ensured that there was no strong difference in licking rate between rewarded and unrewarded licks. To “LEAVE”, mice had to restart running above the threshold for movement for more than 150 ms, and travel a fixed distance on the treadmill (around 16 cm) to reach the other arm. We defined as correct bouts the ones in which mice stopped licking after the states transitioned from “ACTIVE” to “INACTIVE”. Error bouts were ones in which mice stopped licking before the state transition occurred. In this case, mice had to travel double the distance to get back to the arm in “ACTIVE” state. Missed bouts were ones in which mice alternated between arms without any valid lick. These ‘missed bouts’ were excluded from our analysis.

Mouse training. Mice were handled by the experimenter from 3 to 7 days, starting from the beginning of the water restriction and prior to the first training session. At the beginning of the training, mice were acclimatized to the head-fixation and to the arm movement and received liquid reward simply by licking at the lick-port. The position of the lick-ports relative to the snout of the mouse had an important effect on behavioral performances. Thus, to ensure that the position of the lick-ports remained unchanged across experimental sessions, it was carefully adjusted on the first session and calibrated before the beginning of every other session. After mice learned to lick for water reward (typically after one or two sessions), the next sessions consisted of an easier version of the task (with low probability of state transition, typically 5% or 10%, and high probability of reward delivery, 90%), and both arms in "ACTIVE" state. That way, if mice alternated between arms before the states of the sites transitioned, the other arm would still deliver reward and animals would not receive the travel penalty. Occasionally during the early phase of training, manual water delivery was necessary to motivate the mice to lick or to stop running. Alternatively, it was sometimes necessary to gently touch the tail of the animals, such that they started to run and gradually associated running with the movement of the arms. The difficulty of the following sessions was progressively increased by increasing the probability of state transition if the performance improved. Performance improvement was indicated by an increase in the number of bouts and licking rate, and by a decrease in the average time of different events within a bout. Mice were then trained for at least five consecutive days on the final task to ensure a stable behavior before the recording sessions. The behavior was considered stable when consecutive sessions resulted in no significant differences in the distribution of consecutive failures and in the number of consecutive failures as a function of cumulative rewards.

Electrophysiology. Recordings were made using electrode arrays with 374 recording sites (Neuropixels "Phase3A"). The Neuropixels probes were mounted on a custom 3D-printed piece attached to a stereotaxic apparatus (Kopf Instruments). Before each recording session, the shank of the probe was stained with red-fluorescent dye (DiI, ThermoFisher Vybrant V22885) to allow later track localization. Mice were habituated to the recording setup for a few days prior to the first recording session. Prior to the first recording session, mice were briefly anaesthetized with isoflurane and administered a non-steroidal analgesic (Carprofen) before drilling one small craniotomy (1 mm diameter) over the secondary motor cortex. The craniotomy was cleaned with a sterile solution and covered with silicone sealant (Kwik-Sil, World Precision Instruments). Mice were allowed to recover in their home cages for several hours before the recording. After head-fixation, the silicone sealant was removed and the shank of the probe was advanced through the dura and slowly lowered to its final position. The craniotomies and the ground-pin were covered with a sterile cortex buffer. The probe was allowed to settle for 10 min to 20 min before starting recording. Recordings were acquired with SpikeGLX Neural recording system (https://billkarsh.github.io/SpikeGLX/) using the external reference setting and a gain of 500 for the AP band (300 Hz high-pass filter). Recordings were made from either hemisphere. The target location of the probe corresponded to the coordinates of the anterior lateral motor cortex, a region of the secondary motor cortex important for motor planning of licking behavior (*43*). The probe simultaneously traversed the orbitofrontal cortex, directly ventral to the secondary motor cortex. In a subset of recording sessions (4 out of 7), a large portion of the probe tip ended in the olfactory cortex, ventral to the orbitofrontal cortex. We refer to as 'the frontal cortex' only neocortical regions (i.e., secondary motor cortex and orbitofrontal cortex).

Histology and probe localization. After the recording session, mice were deeply anesthetized with Ketamine/Xylazine and perfused with 4% paraformaldehyde. The brain was extracted and fixed for 24 hours in paraformaldehyde at 4 C, and then washed with 1% phosphate-buffered saline. The brain was sectioned at 50 μm, mounted on glass slides, and stained with 4',6-diamidino-2-phenylindole (DAPI). Images were taken at 5x magnifications for each section using a Zeiss AxioImager at two different wavelengths (one for DAPI and one for DiI). To determine the trajectory of the probe and approximate the location of the recording sites, we used SHARP-Track (*44*), an open-source tool for analyzing electrode tracks from slice histology. First, an initial visual guess was made to find the coordinates from the Allen Mouse Brain Atlas (3D Allen CCF, http://download.alleninstitute.org/informatics-archive/current-release/mouse_ccf/annotation/) for each DiI mark along the track by comparing structural aspects of the histological slice with features in the atlas. Once the coordinates were identified, slice images were registered to the atlas using manual input and a line was fitted to the DiI track 3D coordinates. As a result, the atlas labels along the probe track were extracted and aligned to the recording sites based on their location on the shank. Finally, we also used characteristic physiological features to refine the alignment procedure (i.e, clusters of similar spike amplitude across cortical layers, low spike rate between frontal and olfactory cortical boundaries, or LFP signatures in deep olfactory areas).

Analysis and statistics

Pre-processing neural data. Neural data were pre-processed as described previously (*45*). Briefly, the neural data were first automatically spike-sorted with Kilosort2 (https://github.com/MouseLand/Kilosort) using MATLAB (MathWork, Natick, MA, USA). To remove the baseline offset of the extracellular voltage traces, the median activity of each channel was subtracted. Then, to remove artifacts, traces were "common-average referenced" by subtracting the median activity across all channels at each time point. Second, the data was manually curated using an open source neurophysiological data analysis package (Phy:https://github.com/kwikteam/phy). This step consisted in categorizing each cluster of events detected by a particular Kilosort template into a good unit or an artifact. There were several criteria to judge a cluster as noise (non-physiological waveform shape or pattern of activity across channels, spikes with inconsistent waveform shapes within the same cluster, very low-amplitude spikes, and high contamination of the refractory period). Units labeled as artifacts were discarded in further analyses. Additionally, each unit was compared to spatially neighboring units with similar waveforms to determine whether they should be merged, based on cross-correlogram features and/or drift patterns. Units passing all these criteria were labeled as good and considered to reflect the spiking activity of a single neuron. For all analyses, otherwise noted, we averaged for each neuron the number of spikes into bins by considering a 200 ms window centered around each lick. The bin-vectors were then z-scored. Because the interval between each lick was on average around 150 ms there was little overlap between two consecutive bins and each bin typically contained the number of spikes associated with only one lick. Yet, we also tested different window sizes (100 ms and 300 ms centered around the lick or 150 ms after the lick) and the results held (i.e., the different decision variables could still be decoded with high accuracy).

Analysis across individuals. All data analyses were performed with custom-written software using MATLAB (MathWork, Natick, MA, USA). For the large behavioral groups (n = 21 mice or n = 120 sessions), mean and standard deviation (s.d.) of the mean were reported. For the smaller group of recorded mice (n = 7) median and median absolute deviation (m.a.d.) were reported and nonparametric methods were used to compare across conditions.

Predicting choice from decision variables. We used logistic regression (*46*) to estimate how decision variables predicted the choice of the animal (i.e., the probability that the current lick is the last in the bout). Using the Matlab function fitglm (Glmnet for Matlab (2013) Qian, J., Hastie, T., Friedman, J., Tibshirani, R. and Simon, N.; http://www.stanford.edu/~hastie/glmnet_matlab/) with binomial distribution, model fits were performed with decision variables as predictors. We used 5-fold nested cross-validation and elastic net regularization (α = 0.5). To assess a metric of model fit, we calculated the deviance explained (as implemented by the devianceTest function in Matlab). The deviance explained is a global measure of fit that is a generalization of the determination coefficient (r-squared) for generalized linear models. It is calculated as:

Deviance explained = 1 - residual deviance / null deviance.

The residual deviance is defined as twice the difference between the log-likelihoods of the perfect fit (i.e., the saturated model) and the fitted model. The null deviance is the residual deviance of the worse fit (i.e the model that only contains an intercept). The log-likelihood of the fitted model is always smaller than the log-likelihood of the saturated model, and always larger than the log-likelihood of the null model. As a consequence, if the fitted model does better than the null model at predicting choice, the resulting deviance explained should be between 0 and 1. When the fitted model does not predict much better than the null model, the deviance explained is close to zero.

Predicting decision variables from neural population. We used a generalized linear regression model for Poisson response (*47*) to predict each decision variable given the activity of the neural population. Specifically, we predicted the decision variable *A* given the neural activity *x*, by learning a model with parameters, β, such as $A = exp(\beta_0 + \beta x)$. The Poisson regression with log-link is appropriate to model count data like the decision variables studied here. To enforce positivity of the count responses, we shifted all the decision variables to have a minimum value of one. Model fits were performed on each session separately. Model fits were performed using the Matlab version of the open-source glmnet package (https://web.stanford.edu/~hastie/glmnet_matlab/). We employed elastic net regularization with parameter α = 0.5. Additionally, we performed a cross-validation implemented by cvglmnet using the lamda_min option to select the hyper-parameter that minimizes prediction error. To assess the predictive power of the model, we also implemented a nested cross-validation. Specifically, the model coefficients and hyperparameters were sequentially fit using a training set consisting of four-fifths of the data and the prediction was evaluated on the testing set consisting of the remaining one-fifth. The method was implemented until all the data had been used both for training and testing. The deviance explained reported as a metric of the goodness of fit was calculated from the cross-validated results. The final β coefficients were estimated using the full dataset.

Predicting choice from neural population. We used logistic regression (*46*) to estimate how the weighted sum of neural activity (i.e., the neural projections onto the weights that best predict the various decision variables) predicted the probability that the current lick is the last in the bout. The model fit each recording session separately as described above using the glmnet package in Matlab and implementing elastic net regularization with $\alpha = 0.5$ and a nested 5-fold cross validation to estimate the deviance explained.

Model

Integrator framework. We developed a unified theory of integration in the setting of non-sensory decision making tasks. In a wide variety of tasks, animals need to keep track of quickly evolving external quantities. Here, we considered tasks where the feedback that the animal receives is binary (e.g. reward or failure). We considered an integrator given by $x_{t+1} = a_1 \bullet x_t + b_1$, if the attempt is rewarded, and $x_{t+1} = a_0 \bullet x_t + b_0$, otherwise. The parameters of the integrator a_0 and a_1 represent the computations and are bound between zero and one ($a = 1$ for an accumulation, $a = 0$ for a reset). The parameters b_1, b_0 add linearly and can be negative, positive or null.

We consider different scenarios involving a combination of computations but where the optimal solution only involves a one-dimensional integration. For instance, counting tasks can be solved by a linear integration, i.e. $a_0 = a_1 = b_0 = b_1 = 1$, where the integrated value increases by one after each attempt regardless of the outcome. In a two-alternative forced choice and more generally in an n-armed bandit task, each arm would have an integrator that increases with rewards i.e, $a_0 = a_1 = 1$, $b_0 =0$ and $b_1 = 1$, and decays with failures, i.e., $a_0 = a_1 = 1$, $b_0 = -1$ and $b_1 = 0$. Even in cognitively more complex tasks, involving inference over hidden states, such as reversal tasks or foraging under uncertainty, a single integrator is often sufficient. Specifically in the foraging task studied here, the optimal solution is to integrate failures but not rewards, i.e., $a_0 = b_0 = 1$, and $a_1 = b_1 = 0$.

More generally, the model produces sequences that ramp-up with failures (i.e., $a_0 = b_0 = 1$; such as the consecutive failures), and the mirror images that ramp down (i.e., $a_0 = 1$, $b_0 = -1$). Similarly, the model can produce sequences that ramp-up or down with rewards (i.e., $a_1 = 1$, $b_1 = \pm 1$). The model also generates sequences that accumulate one type of event and persist at a constant level with the other type (i.e., $a_x = 1$, $b_x = \pm 1$, $a_Y = 1$, $b_y = 0$), such as the cumulative reward integrator or its mirror image. Finally, many sequences generated by the model (where $a_0 = a_1 =0$) track the outcomes (i.e., reward vs. failure).

There are 36 different values that the parameters of the model can take (a_0 and a_1 could take the values of 0 or 1 and b_0 and b_1 could take the values of -1, 0 or 1). In principle, each of these defines a different model which generates a time-series when fed with sequences of binary action outcomes. The 8 of them for which $b_0 = b_1=0$ are trivial (constant). Of the remaining 28, not all are linearly independent. For instance, the time series generated by the model that computes ‘count’ ($a_0 = a_1 = b_0 = b_1 = 1$) is equal to the sum of the time series generated by the model that accumulates reward and is insensitive to failures ($a_0 = a_1 = 1$; $b_0 = 0$; $b_1 = 1$) and the time series generated by the model that accumulates failures and is insensitive to rewards ($a_0 = a_1= 1$; $b_0 = 1$;

$b_1 = 0$). Thus, the rank of the space of time series is 8 (two dimensions for the linear component ($b_{0,1}$) of the model for each of the four possible combinations of the $a_{0,1}$ parameters, which specify the 'computation' the model is performing). Out of these 8 dimensions, 4 come from models that are less interesting. Two of these are the two 'outcome' time series ($a_0 = a_1 = 0$), which are 'observable'. We also only consider one time series for each of the two integrate-and-reset models, since the value of the linear component associated with the outcome that is reset makes very little difference to the overall shape of the time series. For instance, the time series generated by the two models $a_0 = 1$; $a_1 = 0$; $b_0 = 1$; $b_1 = 0$ and $a_0 = 1$; $a_1 = 0$; $b_0 = 1$; $b_1 = 1$ are linearly independent but almost identical for the type of outcome sequences of interest. The remaining 4 dimensions after these 'trivial' models are removed are spanned by the 4 basis elements that we focus on in the main text (Fig. 4). Finally, the effective dimensionality of the space of time series also depends on the temporal statistics of the outcome sequences. For the particular outcome sequences experienced by the mice (which are a function of the reward and state-transition probabilities) the effective dimensionality was low, which motivated us to focus on particular subsets of outcome sequences in Fig. 4 where the time series generated by the 4 basis elements are clearly distinct

Acknowledgments

We thank Pietro Vertechi for insightful discussions about the project and the model, Joao Morais for help with behavioral training, Davide Reato for support with analyses, and Luca Mazzucato for comments on the manuscript. We also thank Michael Beckert for assistance with the illustrations. This work was supported by an EMBO long-term fellowship (F.C.; ALTF 461-2016) an AXA postdoctoral fellowship (F.C.), the MEXT Grant-in-Aid for Scientific Research (19H05208, 19H05310, 19K06882, M.M.), the Takeda Science Foundation (M.M.), Fundação para a Ciência e a Tecnologia (PTDC/MED_NEU/32068/2017, M.M. Z.F.M.; and LISBOA-01-0145-FEDER-032077, A.R.), the European Research Council Advanced Grant (671251, Z.F.M.), Simons Foundation (SCGB 543011, Z.F.M.), and Champalimaud Foundation (Z.F.M., A.R.). This work was also supported by Portuguese national funds, through FCT - Fundação para a Ciência e a Tecnologia - in the context of the project UIDB/04443/2020 and by the research infrastructure CONGENTO, co-financed by Lisboa Regional Operational Programme (Lisboa2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) and Fundação para a Ciência e Tecnologia (Portugal) under the projects LISBOA-01-0145-FEDER-02217 and LISBOA-01-0145-FEDER-022122.

Author contributions

F.C. and Z.F.M. conceived the project. F.C. and M.M. designed and performed behavioral experiments. F.C. designed and performed electrophysiological experiments. F.C. curated the data. F.C. and A.R. designed and performed the analyses. F.C., A.R and Z.F.M. wrote the manuscript. All authors reviewed and edited the manuscript.

References

1. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. Neuron. **81**, 267–279 (2014).

2. P. Vertechi, E. Lottem, D. Sarra, B. Godinho, I. Treves, T. Quendera, M. N. Oude Lohuis, Z. F. Mainen, Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas. Neuron. **106**, 166-176.e6 (2020).

3. Y. Niv, Learning task-state representations. Nat. Neurosci. **22**, 1544–1553 (2019).

4. M. F. S. Rushworth, T. E. J. Behrens, Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. **11**, 389–397 (2008).

5. L. P. Sugrue, G. S. Corrado, W. T. Newsome, Matching behavior and the representation of value in the parietal cortex. Science. **304**, 1782–1787 (2004).

6. B. W. Brunton, M. M. Botvinick, C. D. Brody, Rats and humans can optimally accumulate evidence for decision-making. Science. **340**, 95–98 (2013).

7. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature. **503**, 78–84 (2013).

8. D. Sussillo, M. M. Churchland, M. T. Kaufman, K. V. Shenoy, A neural network that finds a naturalistic solution for the production of muscle activity. Nat. Neurosci. **18**, 1025–1033 (2015).

9. H. Jaeger, H. Haas, Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. Science. **304**, 78–80 (2004).

10. D. Sussillo, L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks. Neuron. **63**, 544–557 (2009).

11. R. Laje, D. V. Buonomano, Robust timing and motor patterns by taming chaos in recurrent neural networks. Nat. Neurosci. **16**, 925–933 (2013).

12. F. Cazettes, D. Reato, J. P. Morais, A. Renart, Z. F. Mainen, Phasic Activation of Dorsal Raphe Serotonergic Neurons Increases Pupil Size. Curr. Biol. **31**, 192-197.e4 (2021).

13. G. B. M. Mello, S. Soares, J. J. Paton, A Scalable Population Code for Time in the Striatum. Curr. Biol. **25**, 1113–1122 (2015).

14. J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, Ç. Aydın, M. Barbic, T. J. Blanche, V. Bonin, J. Couto, B. Dutta, S. L. Gratiy, D. A. Gutnisky, M. Häusser, B. Karsh, P. Ledochowitsch, C. M. Lopez, C. Mitelut, S. Musa, M. Okun, M. Pachitariu, J. Putzeys, P. D. Rich, C. Rossant, W. Sun, K. Svoboda, M. Carandini, K. D. Harris, C. Koch, J. O'Keefe, T. D. Harris, Fully Integrated Silicon Probes for High-Density Recording of Neural Activity. Nature. **551**, 232–236 (2017).

15. N. J. Powell, A. D. Redish, Representational changes of latent strategies in rat medial prefrontal cortex precede changes in behaviour. Nat. Commun. **7**, 12830 (2016).

16. R. Bartolo, B. B. Averbeck, Prefrontal Cortex Predicts State Switches during Reversal Learning. Neuron. **106**, 1044-1054.e4 (2020).

17. M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, S. Fusi, The importance of mixed selectivity in complex cognitive tasks. Nature. **497**, 585–590 (2013).

18. D. Raposo, M. T. Kaufman, A. K. Churchland, A category-free neural population supports evolving demands during decision-making. Nat. Neurosci. **17**, 1784–1792 (2014).

19. D. Kobak, W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X.-L. Qi, R. Romo, N. Uchida, C. K. Machens, Demixed principal component analysis of neural population data. eLife. **5**, e10989 (2016).

20. A. Wald, Sequential Analysis. (John Wiley & Sons, New York., 1947).

21. J. Drugowitsch, R. Moreno-Bote, A. K. Churchland, M. N. Shadlen, A. Pouget, The Cost of Accumulating Evidence in Perceptual Decision Making. J. Neurosci. **32**, 3612–3628 (2012).

22. J. I. Gold, M. N. Shadlen, Banburismus and the Brain: Decoding the Relationship between Sensory Stimuli, Decisions, and Reward. Neuron. **36**, 299–308 (2002).

23. C. M. Glaze, J. W. Kable, J. I. Gold, Normative evidence accumulation in unpredictable environments. eLife. **4**, e08825 (2015).

24. I. Krajbich, A. Rangel, Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. Proc. Natl. Acad. Sci. **108**, 13852–13857 (2011).

25. T. Yang, M. N. Shadlen, Probabilistic reasoning by neurons. Nature. **447**, 1075–1080 (2007).

26. M. Sarafyazd, M. Jazayeri, Hierarchical reasoning by neural circuits in the frontal cortex. Science. **364** (2019), doi:10.1126/science.aav8911.

27. R. S. Sutton, A. G. Barto, Reinforcement learning: an introduction (MIT Press, Cambridge, Mass, 1998), Adaptive computation and machine learning.

28. L. P. Kaelbling, M. L. Littman, A. R. Cassandra, Planning and acting in partially observable stochastic domains. Artif. Intell. **101**, 99–134 (1998).

29. R. P. N. Rao, Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. Front. Comput. Neurosci. **4** (2010), doi:10.3389/fncom.2010.00146.

30. A. Hermoso-Mendizabal, A. Hyafil, P. E. Rueda-Orozco, S. Jaramillo, D. Robbe, J. de la Rocha, Response outcomes gate the impact of expectations on perceptual decisions. Nat. Commun. **11**, 1057 (2020).

31. Q. Xiong, P. Znamenskiy, A. M. Zador, Selective corticostriatal plasticity during acquisition of an auditory discrimination task. Nature. **521**, 348–351 (2015).

32. J. Drugowitsch, A. G. Mendonça, Z. F. Mainen, A. Pouget, Learning optimal decisions with confidence. Proc. Natl. Acad. Sci. **116**, 24872–24880 (2019).

33. G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, A. Hirose, Recent advances in physical reservoir computing: A review. Neural Netw. **115**, 100–123 (2019).

34. P. Enel, E. Procyk, R. Quilodran, P. F. Dominey, Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. PLOS Comput. Biol. **12**, e1004967 (2016).

35. M. Sigman, S. Dehaene, Parsing a Cognitive Task: A Characterization of the Mind's Bottleneck. PLOS Biol. **3**, e37 (2005).

36. A. Zylberberg, B. Ouellette, M. Sigman, P. R. Roelfsema, Decision Making during the Psychological Refractory Period. Curr. Biol. **22**, 1795–1799 (2012).

37. Y. H. Kang, A. Löffler, D. Jeurissen, A. Zylberberg, D. M. Wolpert, M. N. Shadlen, bioRxiv, in press, doi:10.1101/2020.10.15.341008.

38. P. Cisek, Cortical mechanisms of action selection: the affordance competition hypothesis. Philos. Trans. R. Soc. B Biol. Sci. **362**, 1585–1599 (2007).

39. J. P. Gallivan, L. Logan, D. M. Wolpert, J. R. Flanagan, Parallel specification of competing sensorimotor control policies for alternative action options. Nat. Neurosci. **19**, 320–326 (2016).

40. A. Shenhav, M. A. Straccia, S. Musslick, J. D. Cohen, M. M. Botvinick, Dissociable neural mechanisms track evidence accumulation for selection of attention versus action. Nat. Commun. **9**, 2485 (2018).

41. S. T. Klapp, D. Maslovat, R. J. Jagacinski, The bottleneck of the psychological refractory period effect involves timing of response initiation rather than response selection. Psychon. Bull. Rev. **26**, 29–47 (2019).

42. G. Lopes, N. Bonacchi, J. Frazão, J. P. Neto, B. V. Atallah, S. Soares, L. Moreira, S. Matias, P. M. Itskov, P. A. Correia, R. E. Medina, L. Calcaterra, E. Dreosti, J. J. Paton, A. R. Kampff, Bonsai: an event-based framework for processing and controlling data streams. Front. Neuroinformatics. **9** (2015), doi:10.3389/fninf.2015.00007.

43. N. Li, T.-W. Chen, Z. V. Guo, C. R. Gerfen, K. Svoboda, A motor cortex circuit for motor planning and movement. Nature. **519**, 51–56 (2015).

44. P. Shamash, M. Carandini, K. Harris, N. Steinmetz, A tool for analyzing electrode tracks from slice histology. bioRxiv, 447995 (2018).

45. N. A. Steinmetz, P. Zatka-Haas, M. Carandini, K. D. Harris, Distributed coding of choice, action, and engagement across the mouse brain. Nature. **576**, 266–273 (2019).

46. N. Simon, J. H. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Softw. **39**, 1–13 (2011).

47. J. H. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. **33**, 1–22 (2010).

48. D. M. Green, J. A. Swets, Signal detection theory and psychophysics (New York : Wiley, 1966; https://trove.nla.gov.au/version/12407339).

49. C. E. Feierstein, M. C. Quirk, N. Uchida, D. L. Sosulski, Z. F. Mainen, Representation of Spatial Goals in Rat Orbitofrontal Cortex. Neuron. **51**, 495–507 (2006).

Supplementary Information

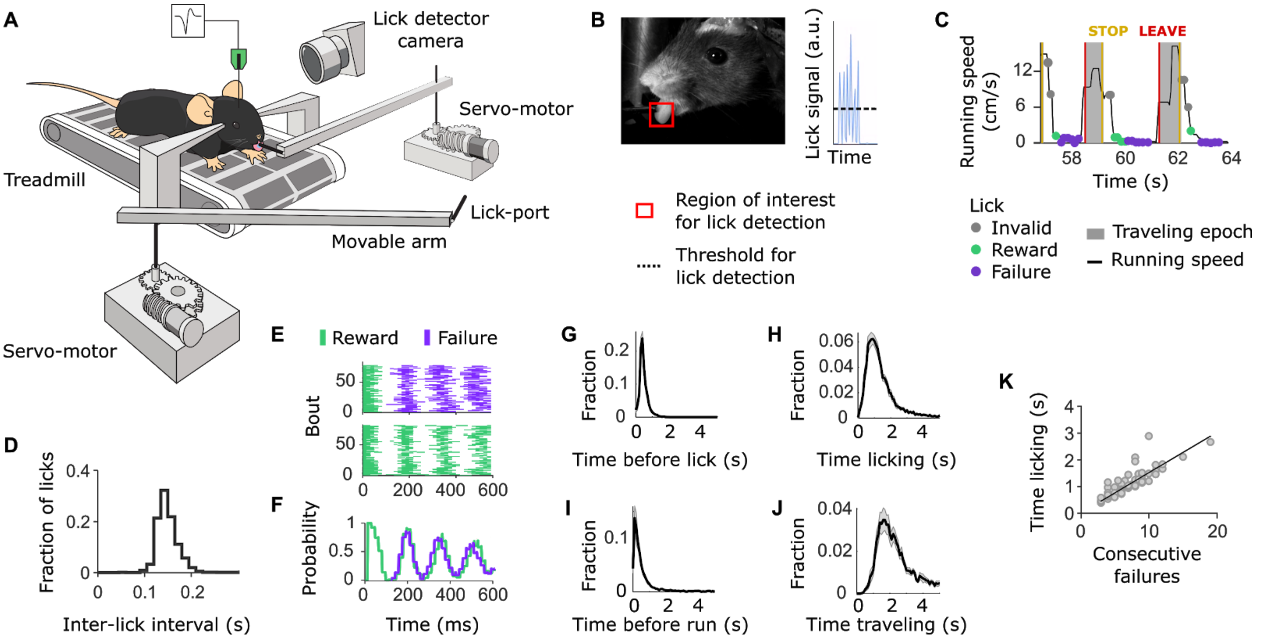

Fig. S1. Task apparatus and behavioral properties. (A) The behavioral apparatus consists of a treadmill, coupled to two servo motors. Rotating the treadmill activates in a closed-loop fashion the movement of the arms via the motors. A mouse placed on the treadmill with its head fixed can lick at the spout from the arm in front. A camera placed on the side of the animal allows on-line video detection of the licks. (B) View from the lick detector camera. A region of interest is defined around the tongue of the animal. To detect the licks a threshold is applied to the signal within the region of interest. (C) The task consists of behavioral bouts and traveling epochs. Within a behavioral bout, the outcomes of the licks are classified into three types: reward, failure and invalid. Rewards and failures occur when the mouse slows down its running speed below an arbitrary threshold after the 'STOP event'. The 'STOP event' is signaled by an auditory tone when an arm comes into place. Any lick above the running threshold is considered as invalid and always unrewarded. The travelling epoch starts after the 'LEAVE event', when the mouse initiates the run. (D, E, F) The licking behavior of the animals is stereotyped. (D) Histogram of the time between each lick. (E) Examples of lick raster of consecutive failures (top) and consecutive rewards (bottom). Licks are aligned at the onset of a rewarded lick and sorted based on the following events. (F) The licking frequency corresponding to the two different examples (series of consecutive rewards in green and series of consecutive failures in purple). (G, H, I, J) Time distributions of different behavioral events (mean ± s.e.m; n = 21 mice). The time spent licking was much greater than the time to initiate licking (between STOP event and first lick) or the time to initiate running (between the last lick and LEAVE event). Notably, engaged mice took less than half a second after the last licks to leave the site in the majority of bouts (Median time to run = 0.46 s). The running time is comparable to the licking time. (K) Monotonic relationship between the number of consecutive failures after the last reward and the time licking after the last reward (each dot represents the means across bouts for each session).

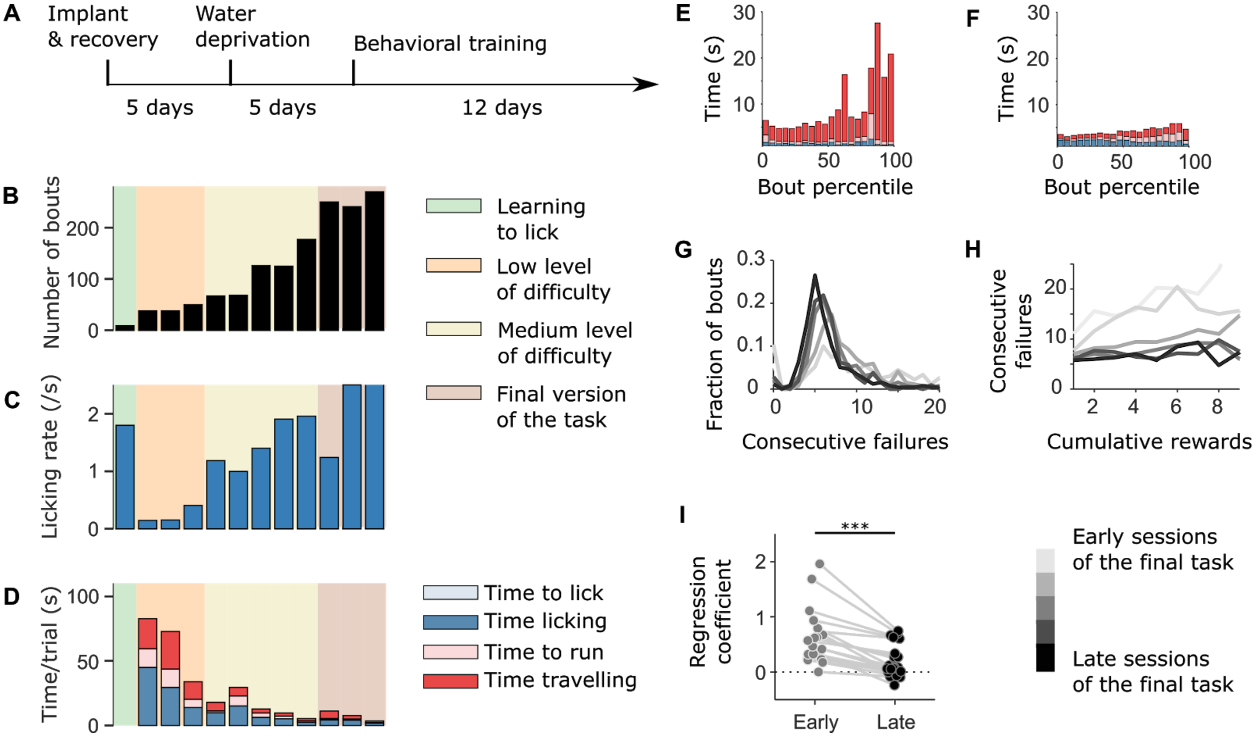

Fig. S2. Mice training. (A) Training timeline. Overall, mice quickly learned the task (from 10 days to a month of training) (B, C, D) Mice went through several phases of training during which the difficulty of the task was progressively increased according to the task performances. Task performances during training were assessed by: (B) the increased number of bouts performed across training days, (C) the increased licking rate within the session, and (D) the decreased time to complete the behavioral bouts and the traveling epochs. (E, F) Examples of the effects of time and satiety within a session. For some mice, task performances tended to degrade throughout the session (E), while the performances of other mice were barely affected by time and satiety (F). (G, H) Example summary statistics for one animal across training sessions. Several sessions (light grey early sessions to dark grey late sessions) of the final protocol were necessary to reach a stable behavior. The distribution of the total number of consecutive failures (G) and the relationship between consecutive failures and cumulative rewards (H) shifted across training days before they stabilized. (I) Regression coefficients indicating the slope of the curves in (H) for early training sessions and late training sessions for each animal. The smaller the coefficient, the weaker the relationship between consecutive failures and cumulative rewards. Eventually, many mice learned to largely ignore the number of rewards (stars represent the significance p < 0.001 of a paired t-test).

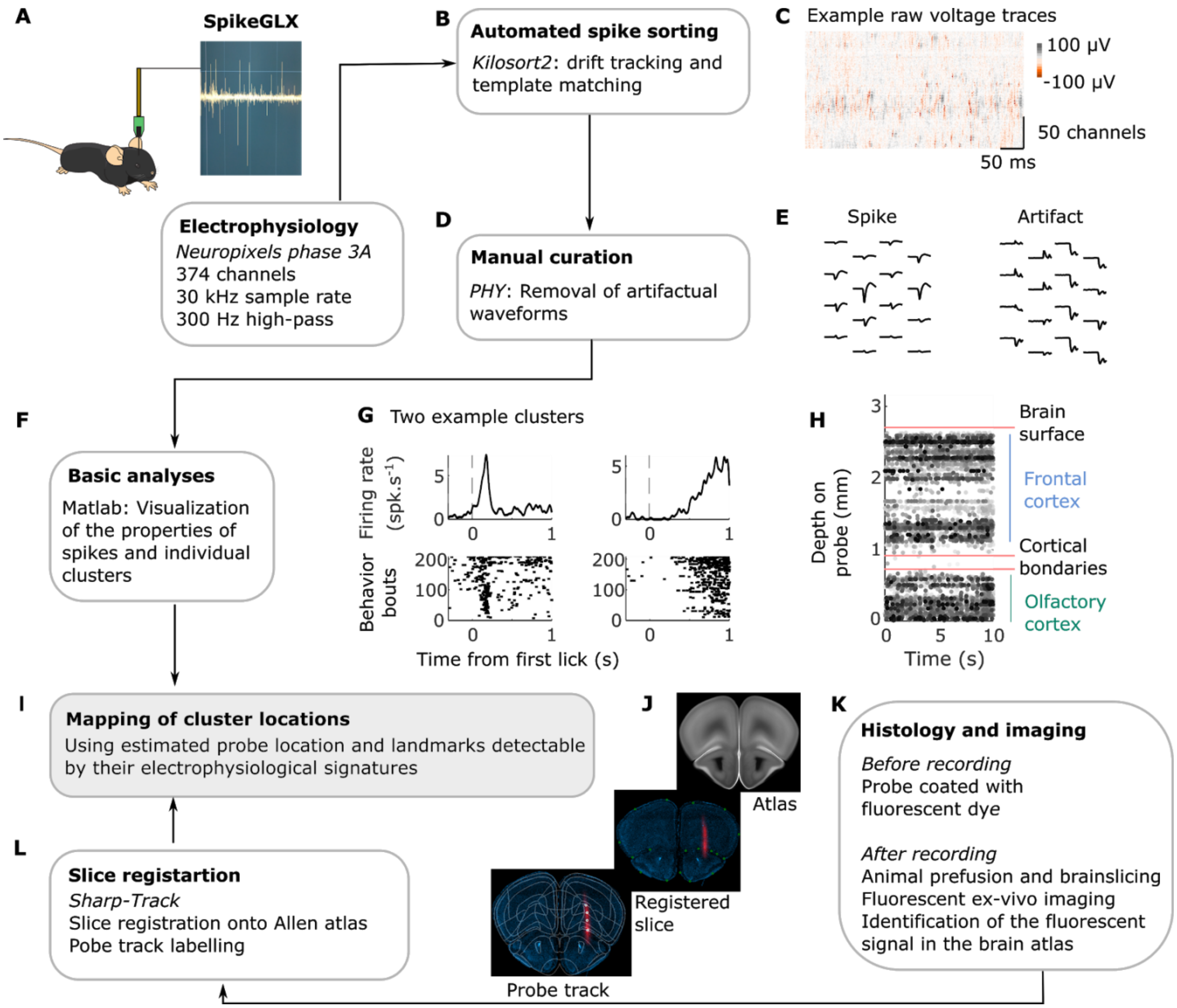

Fig. S3. Pipeline for extracellular electrophysiology, data processing and cluster mapping. (A) Data collection from the Neuropixels probe. (B) Kilosort2 is used to automatically match spike templates to raw data. (C) Example of voltage data input to Kilosort2. Prior to the automatic sorting, the raw data is pre-processed with offset subtraction, median subtraction, and whitening steps. (D) Manual quality control is done on the outputs of Kilosort2 using PHY to remove units with non-physiological waveforms (E), contaminated refractory periods, low amplitude (less than 50 µV) or low spiking units (less than 0.5 spike·s^{-1}). (F) For further quality control, visualization of peri-event spike histograms (G, top; examples histogram aligned to first lick) or scatter plots (G, bottom; example scatter plot aligned to first lick) of single neurons are made with custom-written script in Matlab. (H, I) Example scatter plot of all neurons recorded simultaneously along the shank of the probe. This visualization helps delimitate landmarks based on electrophysiological signatures to map cluster locations. (J, K, L) Landmarks derived from electrophysiological responses are validated with estimates from histology using an open-source software (Sharp-Track).

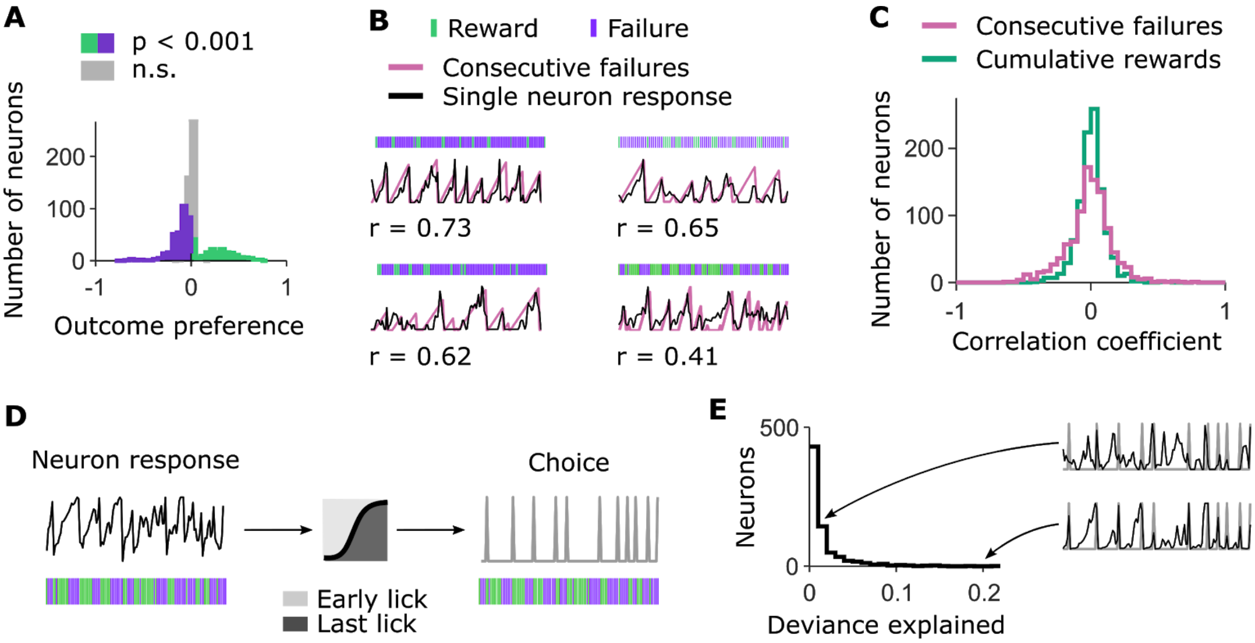

Fig. S4. Single-neuron correlates of decision variables and choices. (A) Histogram of outcome selectivity of all neurons recorded in the frontal cortex. We used receiver operator characteristic (ROC) analysis to assign a preference index to each neuron (*48*). In brief, an ideal observer measures how well the modulation of neuronal firing can classify the outcome (reward or failure) on a trial-by-trial basis. We derived the outcome preference from the area under the ROC curve as defined in Feierstein et al. (2006) (*49*): $PREF_{R,F} = 2[ROC_{AREA}(f_R, f_F) - 0.5]$, where f_R and f_F are the firing rate distributions for trials where outcomes are reward and failure respectively. This measure ranges from −1 to 1, where −1 indicates preference for F (failure), 1 means preference for R (reward) and 0 represents no selectivity. The statistical significance of the preference index ($p < 0.001$, one-sided) was assessed via bootstrapping (1000 iterations). Color bars indicate neurons where the index was significantly different from 0. (B) Example neurons with ramping activity (balck) that correlated well on a bout-by-bout basis with the trajectory of the consecutive failures (pink). The Pearson correlation coefficient between neural activity and the decision variable trajectory (r) is indicated below each trace. (C) Histogram of correlation coefficient between neural response and the trajectories of the integrators for all neurons. Overall, few neurons were highly correlated with the integrator trajectories, especially with the cumulative reward integrator. (D) Logistic regression was used to quantify how well single neuron activity predicted the probability that a given lick was the last of the bout. (E) Histogram of deviance explained from the logistic fits. The large majority of single neurons did not predict well the choice of the animal.

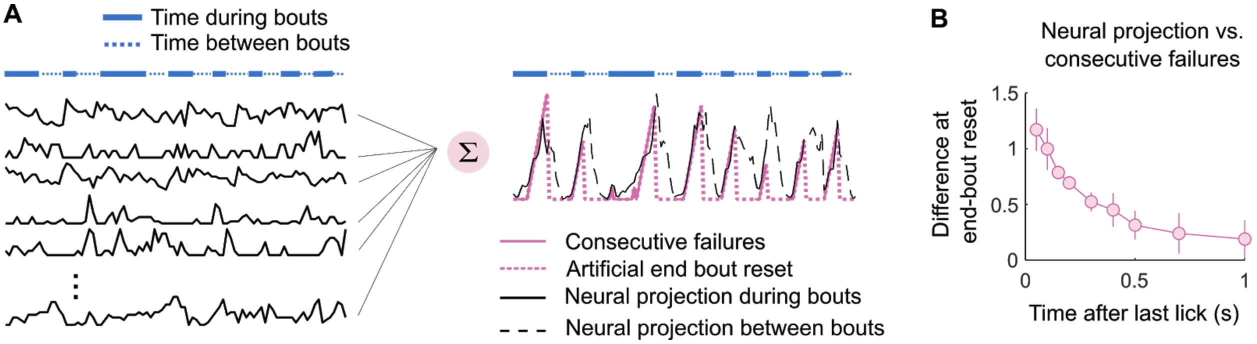

Fig. S5. Time constant of reset at the end of the bout in the frontal cortex. (A) Example consecutive failures (pink) and neural projections (black right) of the neural activity (left, example neural traces) including the activity during 2 s after the end of each bout (dashed-line). The projection of the neural activity on the decoding weights for the consecutive failure slowly ramps down until the beginning of the next bout. (B) To quantify the time constant of the reset at the end of the bout, the consecutive failures with an additional reset at the end of the bout was decoded from the neural activity. We considered the decoding projection at different times after the end of the last lick of bout ‘n’ and before the start of bout ‘n+1’ and plotted the difference between the number of the consecutive failures (dashed pink) and the neural projection (dashed black) at the end of each bout across recording sessions (median ± m.a.d..; n = 7) as a function of the time after the last lick. The neural activity can reset at the end of the bouts with a time constant of around 200 ms.

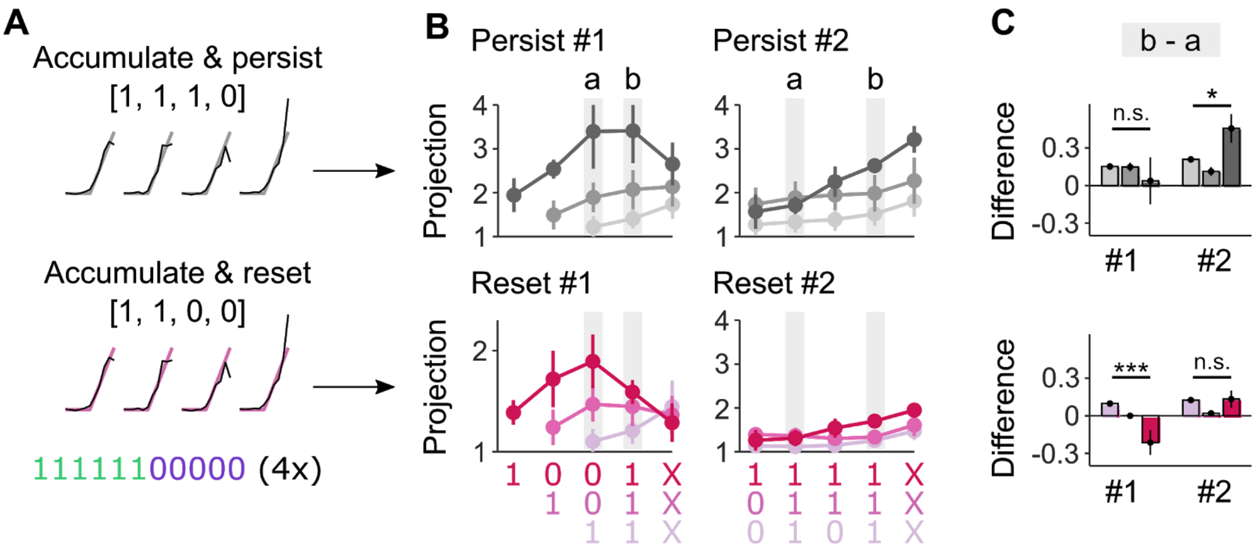

Fig. S6. The two basis elements for 'accumulation of failures and persistence with reward' and 'accumulation of failures and reward reset' are independently represented. (A) The basis element that accumulates failures and persists with rewards (grey) is similar to the consecutive failures (pink) for all sequences where rewards precede failures. Black traces are the neural projections on the decoding weights. (B) Example projection for one recording session (median ± m.a.d. across similar bouts). Selecting sequences of events with interleaved reward and failure (left) or beginning with a failure (right) decorrelates the time series from the accumulation of failures and persistence (top) and the accumulation of failures and reset (i.e., consecutive failures; bottom). (C) Summary across recordings (mean ± s.d., n = 7) of the difference between the projected neural activity at events 'a' and 'b' (highlighted in (B)). This analysis reveals that the neural activity of the frontal cortex can simultaneously persist (top right: one-way ANOVA P(F>7.41) <0.0049, df = 2) and reset (bottom left: one-way ANOVA P(F>17.04) <0.0002, df = 2) with rewards.

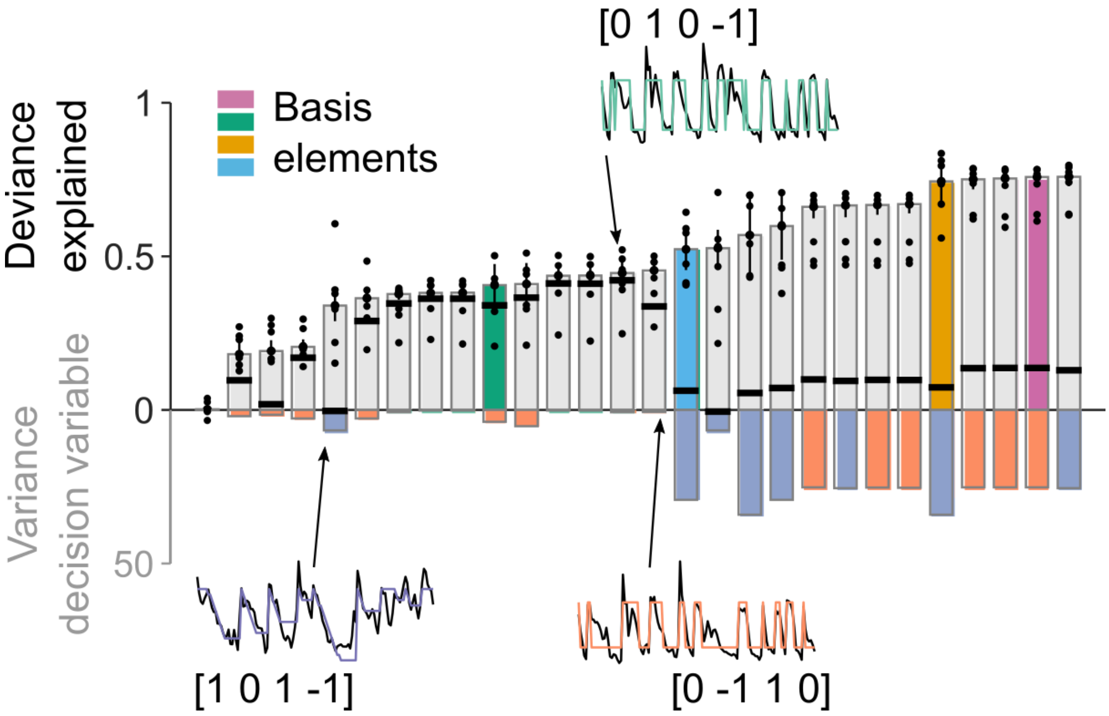

Fig. S7. A reservoir of decision variables decoded from the frontal cortex. The decoding quality of each time series from the repertoire (upward bar is the median across recordings, dots represent single recording sessions, bar is the m.a.d. across sessions), including the basis elements (green, blue, yellow, pink), increased with the variance of the time series (downward bar, orange for time series with reset and purple for time series without reset). Black lines indicate the averaged performance estimated from randomly shuffling of rewards and failures to test explicitly the contribution of observable events to these decision representations (100 shuffles per recording of the population activity among time bins for which the outcome of each lick was the same). Example time series from three different classes of computations (i.e., accumulation, reset and outcome; normalized color traces) and the projections of population activity on the decoding weights for these time series (black traces).

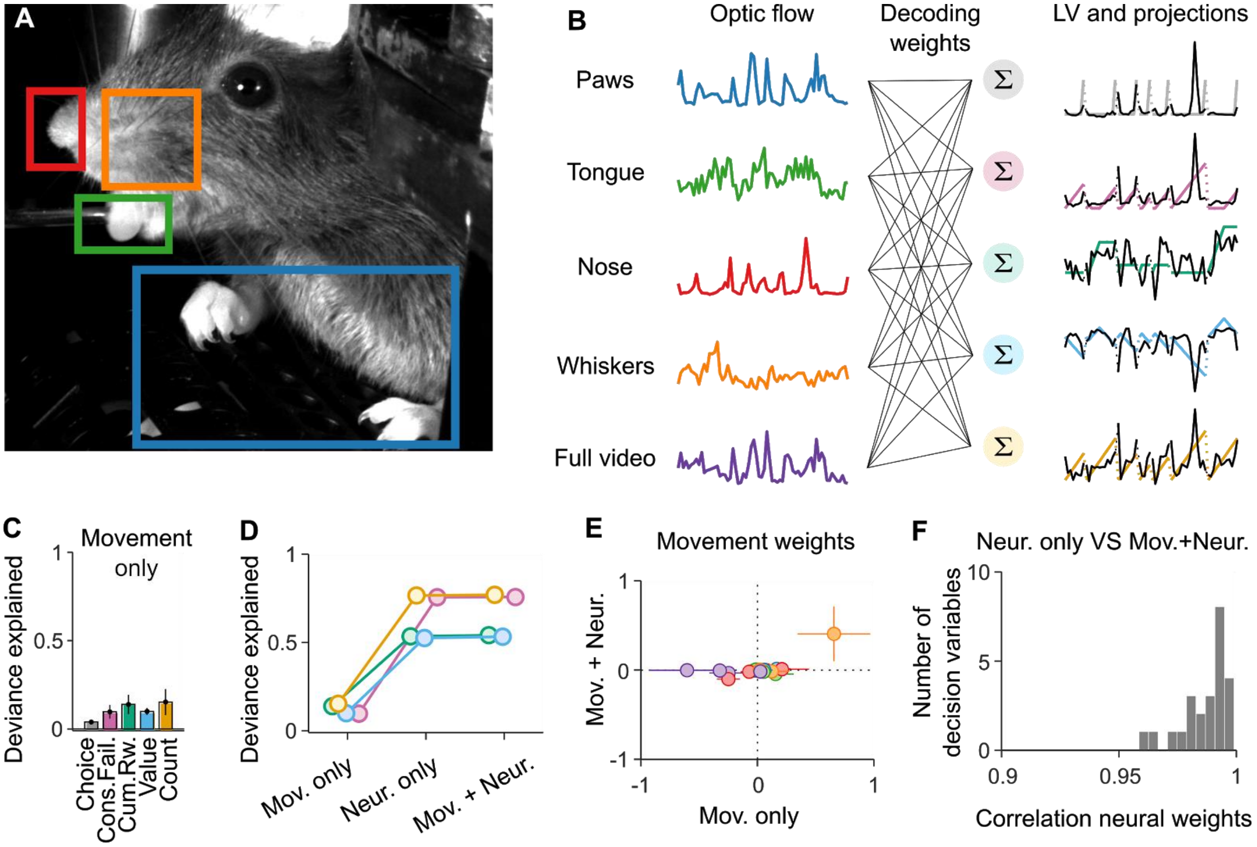

Fig. S8. Predicting decision variables from movements. (A) The optic flow of the video corresponding to each recording session was computed using the Lucas–Kanade method (Matlab function: opticalFlowLK). The optic flow was then extracted from four regions of interest (ROIs) around the forelimbs (blue), the tongue (green), the nose (red) and the whiskers (orange). (B) The average magnitude of the optic flow of the four ROIs and of the entire image were used as predictors in GLMs to explain the choice and the different decision variables in movement-only models (same method as in Fig.3; color traces on the right: grey = choice; pink = consecutive failures; green = cumulative rewards; blue = negative value; yellow = count). (C) The performance of the regressions across all recording sessions (median ± m.a.d..; n = 7) using only the movements as predictors (as exemplified in (B)). (D) The performance of the regressions across all recording sessions (median; n = 7) for models with only movements as predictors (like in (b,c)), models with only neural activities as predictors (line in Fig.3), and models with movement and neurons as predictors. The colors represent the decoding performances for different decision variables. The movements-only models performed significantly worse than the neurons-only models. Additionally, adding the movements to the decoders with neurons (Mov. + Neur.) does not further improve the performances, suggesting that the information provided by the movements is also contained in the neural activity. (E) The betas weights applied to each movement predictor (median ± m.a.d..; n = 7) for the decoding of different decision variables with movement-only models (abscissa) and models including movements and neurons as predictors (ordinate). When neurons are included as predictors (ordinate), the weights applied to

the movements predictors are overall close to zeros, suggesting that the movements provide no further information in explaining the decision variables. (F) The Pearson correlation coefficients between the neural weights vectors of the neurons-only models and the models including neurons and movement as predictors. The coefficients were greater than 0.96 for all decision variables across all recordings, suggesting that adding the movements as predictors in the decoder does not affect the information provided by the neurons.