Preprint of an article accepted for publication in the *Journal of Cognitive Neuroscience*
(submitted on 6 Apr 2015; revised on 20 July and 4 Sept., accepted on 11 September 2015)

Prefrontal goal-codes emerge as latent states in probabilistic value learning

Ivilin Stoianov[1,3], Aldo Genovesio[2], and Giovanni Pezzulo[1*]

[1] Institute of Cognitive Sciences and Technologies, National Research Council
Via S. Martino della Battaglia, 44, Rome 00185, Italy
[2] Dipartimento di Fisiologia Umana e Farmacologia, University La Sapienza
Rome 00185, Italy
[3] Laboratoire de Psychologie Cognitive, CNRS and Aix-Marseille University,
Place Victor Hugo 3, Marseille 13331, France

Email: ivilin.stoianov@gmail.com aldo.genovesio@uniroma1.it (*)giovanni.pezzulo@istc.cnr.it

Keywords: prefrontal cortex, cognitive control, reward learning, neurocomputational model, mutual information

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements: This study was supported by the European Commission's Seventh Framework Programme (grant no. FP7 270108 to G.P. and PIEF-GA-2013-622882 to I.S.) and the Human Frontier Science Program (grant no. RGY0088/2014 to GP). The GEFORCE Titan used for this research was donated by the NVIDIA Corporation.

Author contributions: G.P. and I.S. conceived the experiments. A.G., G.P., and I.S. discussed the results and wrote the paper. I.S. wrote and ran the code and analyzed data.

(*) Corresponding author: Giovanni Pezzulo
Institute of Cognitive Sciences and Technologies
National Research Council
Via S. Martino della Battaglia 44, 00185 Rome, Italy
Telephone: +39 06 44595206; Fax: +39 06 44595243.
E-mail: giovanni.pezzulo@istc.cnr.it

ABSTRACT

The prefrontal cortex (PFC) supports goal-directed actions and exerts cognitive control over behavior but the underlying coding and mechanism are heavily debated. We present evidence for the role of goal-coding in the PFC from two converging perspectives: computational modeling and neuronal-level analysis of monkey data. We show that neural representations of prospective goals emerge by combining a categorization process that extracts relevant behavioral abstractions from the input data and a reward-driven process that selects candidate categories depending on their adaptive value; both forms of learning have a plausible neural implementation in the PFC. Our analyses demonstrate a fundamental principle: goal-coding represents an efficient solution to cognitive control problems, analogous to efficient coding principles in other (e.g. visual) brain areas. The novel analytical-computational approach is of general interest since it applies to a variety of neurophysiological studies.

INTRODUCTION

Flexible cognitive control is fundamental to our everyday activities. It relies on the ability to efficiently learn to extract and fulfill goals, different from habitual decisions that build on stereotyped responses (Dolan and Dayan, 2013). The prefrontal cortex (PFC) supports goal-directed behavior by biasing response selection based on contextual information, goals, and other task-relevant information or task-sets (Koechlin et al., 2003; Koechlin and Hyafil, 2007; Frank and Badre, 2012; Genovesio et al., 2006; Miller and Cohen, 2001; Monsell, 2003; Passingham and Wise, 2012; Reverberi et al., 2012). The neural codes and mechanisms supporting these PFC abilities remain elusive, with contrasting proposals that include mixed selectivity for a large basis of task-related properties (Rigotti et al., 2013) and representations of prospective behavioral goals (Genovesio et al., 2012; Yamagata et al., 2012).

To disentangle these alternatives, here we simulated three tasks previously used to study monkey prefrontal function (Genovesio et al., 2012): a *duration-discrimination*, a *distance-discrimination* and a *match-to-sample* task. To simulate these tasks, we used a probabilistic computational model that fuses unsupervised and value-driven learning. In particular, we used an approximate nonparametric probabilistic category learning method (Anderson, 1991; Sanborn et al., 2010) to infer from experience a set of candidate categories that guide stimulus-action-value transitions, and a reward-sensitive process to select the actual category to be used for action control at any given trial (Collins and Koechlin, 2012; Collins and Frank, 2013).

Like in the monkey studies reported in Genovesio et al. (2012), neural representations of prospective goals – or *goal-codes* - emerged in the model as latent statistical categories grouping noisy stimulus-action-value contingencies in optimal ways. The analysis of the model behavior demonstrates that the emerged *goal-codes* afforded efficient learning and action selection. To assess if the *goal-codes* replicate the coding properties of PFC neurons in the Genovesio et al., (2012) study – that is, an advance representation of the identity of the to-be-selected target stimulus – we compared the monkey PFC data and the latent states learned with the computational model using an information-theoretic approach based on (conditional) mutual information, which permit assessing their specific coding properties while excluding confounds. The analyses revealed that the *goal*-codes emerged in the model replicate with high accuracy key properties of PFC neurons and the goal information is not confounded with other characteristics of stimuli such as their color or magnitude. The results thus provide a novel mechanistic explanation of how the PFC exerts cognitive control by learning prospective goal-codes. Furthermore, our results integrate two influential streams of research on PFC functioning that focus on behavioral control (Miller and Cohen, 2001; Passingham and Wise, 2012) and category learning (Seger and Miller, 2010), respectively, showing that these are complementary processes within a non-parametric probabilistic learning system. In a broader perspective, our study points to hierarchical probabilistic inference as a general framework to understand prefrontal function (Monsell, 2003; Doya et al., 2007; Friston, 2010; Friston et al., 2013; Donoso et al., 2014).

METHODS

Recap of the experimental procedure of the target monkey experiment

Our testbed is a set of three monkey tasks (*duration-discrimination*, *distance-discrimination* and *match-to-sample*) where goals were implicitly defined by stimuli magnitudes and colors/shapes (Genovesio et al., 2012). In the original monkey experiment, trials began with the sequential presentation of visual context stimuli S1 and S2 that were either a red square or a blue circle (hereon we will refer only to their color). Then, each of two target stimuli coding the identity (color and shape) of S1 and S2 appeared on a video monitor, randomly to the left or to the right of the screen center (Figure 1a). The monkey's task was to touch a switch below the stimulus that previously either: lasted longer (*duration-discrimination task*), appeared farther from screen center (*distance-discrimination task*), or appeared twice in the trial (*match-to-sample task*). In each trial of the duration-discrimination task the duration and the identity of the stimuli varied, but not their distance. In the distance-discrimination task the distance and the identity of the stimuli varied, but not their duration. In the match-to-sample task the duration of the stimuli varied but not their identity or distance. (Note that this is a peculiar match-to-sample task, because the same sample is presented twice, not once as usual, in order to have the same number of stimuli presentations as in the duration- and distance-discrimination tasks.) In the duration- and match-to-sample task, the stimuli appeared at screen center and lasted 200 – 1.200 ms, varying in steps of 200 ms (6 levels in totals). In the distance task, the stimuli lasted 1000 ms each and the distance from the screen center varied from 1.6 to 9.4 visual degrees, in steps of 1.6 degrees (in total 6 levels). S1 and S2 had equal probability of either lasting longer in the duration and matching to sample task, or being father from screen center in the distance task. The period between S2 and the presentation of the targets for the response lasted 400 or 800ms.

During the training phase, monkeys learned the three tasks in a sequence using a block design, which included first the duration task, then the distance task, and finally the match-to-sample task, with occasional presentation of blocks containing the previously learned tasks, until adequate performance. When monkeys had average response accuracy of 80% or more in the duration- and distance-tasks and in the match-to-sample task the recording period started. In this test period, the monkeys performed the three tasks in blocks with pseudo-random task-order, during which dorsolateral and caudal PFC neural-cell activity was registered (with means of n=192 trials for duration, n=151 trials for distance, and n=92 trials for match-to-sample task, respectively).

The average neural activity of interest was calculated in the 80-400 ms period after it could be discriminated which was the stimulus with the greatest magnitude; during this period, the PFC neurons were found to carry prospectively the information on the goal identity; see Genovesio et al. (2012) for other details.

Computational modeling methods:

We simulated the experiments using a *probabilistic generative model* that learns to predict reinforcement based on noisy sensory information - context (stimuli S1 and S2 with noisy magnitude and color-identity properties) and target (the color-identities of S1 and S2) - actions (press S1 or S2), and latent states or categories inferred during learning (Figure 1c). We assumed that the latent categories correspond to a population of PFC neurons, where individual neurons have specific preferences (e.g., for a response) and compete for selection. The conditional probability that a given category is selected in a given context corresponds instead to the strength of connectivity between

neurons in the latent category population and input neurons encoding context information (Pouget et al., 2013). Thus, a given context conditionally drives the selection of the latent category most-strongly related with it, which (together with the target stimulus) determines action selection, i.e. the selection of the action with the higher probability to reinforcement.

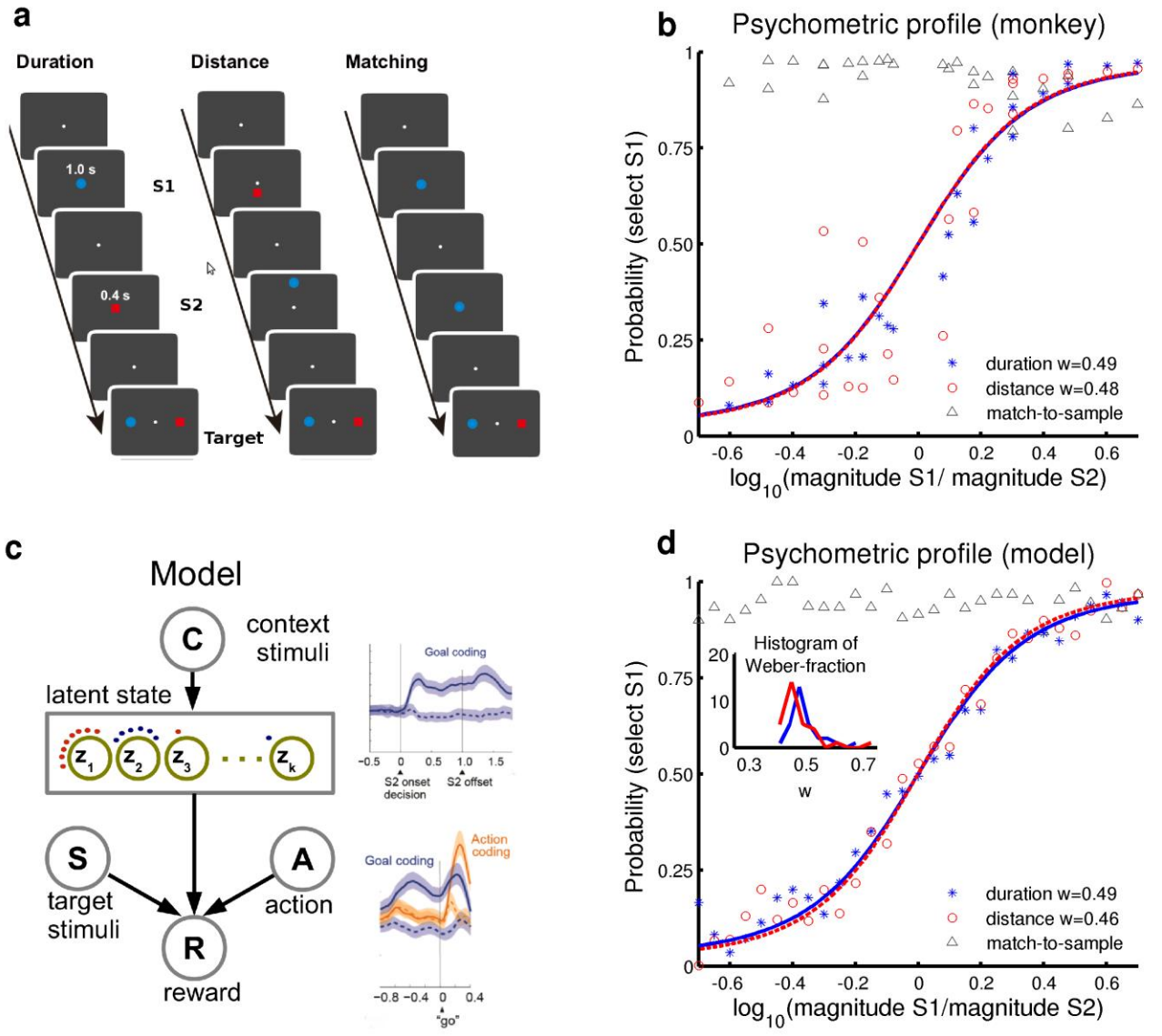

Figure 1: Behavioral analysis. (**a**) Experimental paradigm; details in **Methods**. (**b**) Monkey: *Psychometric profile* of the probability to choose the identity of stimulus S1. For the duration- and distance-discrimination tasks, the probability is plotted as a function of the ratio between S1 and S2 magnitudes, which on a log scale is a sigmoid (Stoianov and Zorzi, 2012) summarized by a variability coefficient, a task-specific Weber fraction *w* (Pica et al., 2004; Stoianov and Zorzi, 2012). For the match-to-sample task, the psychometric profile plots the monkeys' response accuracy as a function of the duration ratio between the two stimuli, displaying independence from it. Stars, circles, and triangles represent data points of the duration, distance, and match task, respectively. Blue, red and gray lines show corresponding sigmoid fits. (**c**-**d**) Model. (**c**) *Bayesian network* representing probabilistic relations between sensory stimuli, latent states (categories), available actions, and reward. *Insets*: top right, population activity averages of the goal cells aligned on the decision point; bottom right, population activity averages of the goal cells and of the response (right or left target location) cells. Solid line, preferred goal (blue) or response (orange), dashed line anti-preferred goal (blue) or response (orange). (**d**) Average psychometric profile of the model, using the same format as in (b).

The model includes a *context* variable that jointly encodes the perceived properties of the stimuli S1 and S2: distance (6 magnitude levels, from closer to farther from the center), duration (6 magnitude levels, from shorter to longer), and color (2 levels: red and blue). The four magnitudes inputs (i.e., the duration and distance of both S1 and S2) were encoded noisily: continuous Gaussian noise was added to each of the four inputs and the resulting continuous values were rounded to obtain their final values (6 magnitude levels each). A fully orthogonal coding of the context variable would require (6*6*2)^2 levels, but since in our tasks only one property (either duration or distance) varies in magnitude in a given trial, we could use a compact coding with only 288 levels[1]. The model includes also a *target* variable that encodes the properties of response-triggering stimuli: color (2 levels) and position (2 levels: left and right). The target variable was orthogonally coded (4 levels). The model finally includes a *response* variable orthogonally coded to represent the two possible actions of the monkey (left and right response), and a *reward* variable orthogonally coded to represent two possible outcomes (rewarded and not rewarded).

Description of the computational model: The model uses a *Bayesian reinforcement learning* scheme to update the conditioned probabilities based on the number of successes and failures. The update considers the (Bernoulli distribution of) predicted reinforcement and the actual action outcomes. In parallel, an approximate non-parametric learning method, *Dirichlet* non-parametric mixture process, shapes the clustering of the contexts into latent categories according to their utility in obtaining reinforcement. First, observed contexts recruit existing clusters according to their overall selection frequency (i.e., popularity) or engage novel ones with a small probability, favoring compact clustering (Gershman and Blei, 2012; Donoso et al., 2014). Second, the conditional probabilities between contexts and categories are scaled depending on the observed stimulus-action-value contingencies; in other words, the probability of reinforcement contributes to the shaping and selection of the categories, see Equation 2 below. This approximate learning method gradually shapes on the acquisition of latent category units that permit to "model" or "explain" the observed stimulus-action-value contingencies and afford efficient reward acquisition (Collins and Koechlin, 2012; Collins and Frank, 2013). The specific method we adopted for the Dirichlet non-parametric mixture process is a local maximum a posteriori (MAP) inference, developed by Sanborn, Griffiths, & Navarro (2010) for category learning and later applied in the domain of reinforcement learning (Collins & Frank, 2013); this non-iterative, local procedure is more biologically realistic compared to alternative (e.g., non local) methods.

In our target monkey task, monkeys had to infer the correct "rule" or "goal" (say, red-goal) based on the sequential presentation of two visual stimuli (a red square and a blue circle) and then select a response (press a button) when the two stimuli were successively presented together, in a randomized (left or right) position. In analogy with this situation, in our model the two earlier presented stimuli S1 and S1 correspond to *context stimuli* c_t that are categorized into latent states, whereas *targets* correspond to response-triggering stimuli s_t (Figure 1c). Importantly, the perceived context stimuli c are clustered into latent states (categories) z according to their utility to obtain

[1] To obtain the *context* variable, the magnitude properties and the color properties of S1 and S2 were combined, using mixed multiplicative-additive coding. First, for each property, duration and distance, the magnitudes of S1 and S2 were multiplicatively combined in an index with 6*6=36 levels. Second, the two (duration- and distance-coding) indexes were additively combined in an overall magnitude index with 36+36=72 levels; the rationale for the additive combination is that, in our tasks, only one among duration or distance varied. Third, the color properties of S1 and S2 were multiplicatively combined in an index with 2*2=4 levels. Finally, the overall representation of the context input was built by multiplicatively combining the magnitude and the color properties in an index of 72*4=288 levels.

reinforcement. The clustering is defined by a probability distribution $P(z|c)$ and initialized with a nonparametric probabilistic approach: a Dirichlet mixture process also known as *Chinese Restaurant Process* (Gershman and Blei, 2012). According to this popular metaphor, clusters, or categories, correspond to restaurant tables and contexts to customers. A newly experienced context (new customer) c_{n+1} is assigned to a new cluster (empty table; new category) z_{new} with a small probability $P(z_{new} \mid c_{n+1}) = \alpha/A$ (controlled by *concentration parameter* α, here α=2) or to an old cluster (occupied table; old category) z_i according to a measure of its popularity $P(z_i \mid c_{n+1}) = \Sigma_j P(z_i \mid c_j) / A$ (table occupancy; category *priors*) across all contexts, where A is a normalizing factor: $A = \alpha + \Sigma_{i,j} P(z_i \mid c_j)$ that essentially counts the number of experienced contexts. Thus, the perceived context c_t evokes the most probable context-specific category (*category inference*) $z_t = argmax_i P(z_i \mid c_t)$ that in turn conditions the successive action selection process and critically to the organization of the category structure, the belief in this assignment $P(z_i \mid c_t)$ is scaled with the probability of the current reinforcement outcome caused by this choice (*category learning*) (Sanborn et al., 2010; Collins and Frank, 2013); see formula (2) below.

The value function defining the reinforcement contingencies is implemented as a binomial probability distribution of reinforcement (rewarded: *r=1*, not rewarded: *r=0*) and conditioned on the target stimulus, action, and latent category (that is in turn conditioned on the context stimulus, see formula (1) and Figure 1C. The distribution is parameterized by a Beta-conjugate distribution, which simply counts the number of successes and failures to obtain reinforcement and affords standard Bayesian learning (3a,b).

$$P(r_t|s_t, a_t, c_t) = \sum_i P(r_t|s_t, a_t, z_i)\, P(z_i|c_t) \qquad (1)$$

When conditioned on the currently selected latent category z_i, target stimulus s_t, and desired reward outcome *r=1*, the probabilistic value function allows inferring the action that most likely brings to that outcome (*action inference*). The chosen action a_t brings to an actual reward outcome r_t.

Experiencing a given trial is followed by an update of the category- and reward- distributions, i.e., learning. The posterior category distribution is updated exploiting an approximate form of Bayesian category learning (Anderson, 1991), accounting here for the observed stimuli-action-value contingencies (Collins and Frank, 2013) (category *learning*).

$$P_{t+1}(z_i|c_t) = \frac{P(r_t|s_t, a_t, z_i)\, P(z_i|c_t)}{\sum_j P(r_t|s_t, a_t, z_j)\, P(z_j|c_t)} \qquad (2)$$

To calculate the posterior of the reward function, we first update the beta-conjugate prior of the conditioned binomial distribution, accounting for the observed reinforcement:

$$n_{t+1}^{r=0}(s_t, a_t, z_t) = n_t^{r=0}(s_t, a_t, z_t) + (1 - r_t) \qquad (3a)$$

$$n_{t+1}^{r=1}(s_t, a_t, z_t) = n_t^{r=1}(s_t, a_t, z_t) + r_t \qquad (3b)$$

Finally, we calculate the posterior conditioned reinforcement distribution P_{t+1} (r=0, r=1) by normalizing $n_{t+1}^{r=1}$ and $n_{t+1}^{r=0}$ to sum to one (*conditional value learning*).

Method for calculating noise in the model stimuli and responses: The perception of sensory magnitudes such as time and distance is intrinsically noisy (e.g., Tudusciuc & Nieder, 2007), such that the perceived magnitudes vary from trial to trial and the variability w_n of the perceived magnitudes scales with the magnitudes according to the Weber-Fechner law (Gibbon, 1977; Dehaene, 2003 Whalen et al., 1999). The noise causes considerable response variability and hinders the learning of magnitude-dependent tasks because of reduced consistency of the experienced stimulus-action-value contingencies. In order to perform a realistic simulation of the monkey experiments in which magnitude comparison plays a crucial role in the duration and distance tasks, we needed to provide our prefrontal model with input stimuli having an adequate level of perceptual noise. We approached this issue by adopting a typical point-wise internal noisy magnitude representation, so-called *mental number line*, characterized by Gaussian noise $w_n = nw_0$ that scales with the magnitude n and which is parameterized by a variability coefficient w_0 (e.g., Whalen et al., 1999). We then conducted a re-analysis of the monkeys' behavior in the original experiment (Genovesio et al., 2012), in order to adequately estimate the variability coefficient w_0 of the perceived relevant (duration and distance) properties.

The first step of the analysis is the calculation of magnitude discriminability coefficients at the behavioral level, the so-called Weber fraction w (Figure 1b). They were obtained by using the monkey responses in the post-learning test period to build psychometric response profiles for each of the three tasks (Figure 1b). For the duration- and distance-discrimination tasks, the profiles display the probability of selecting the identity of stimulus S1 as a function of the log-ratio between the magnitude properties of S1 and S2. Consistent with what reported in the literature, on a log-scale the profile of our magnitude-comparison tasks is a symmetric sigmoidal curve (see e.g., Stoianov and Zorzi, 2012). Essentially, the larger the absolute value of the log-ratio between the compared magnitudes, the larger the probability to select the correct stimulus; or in other words, the more different the magnitudes are, the greater are the odds for a correct response. The profiles are summarized by a magnitude discriminability coefficient, a Weber fraction w that describes the slope of the sigmoid: the smaller the w, the more vertical is the slope and more precise is the response (Pica et al., 2004). For the match-to-sample task, the psychometric profile of Figure 1b plots the monkeys' response accuracy as a function of the duration ratio between the two stimuli, which was manipulated in the task, but irrelevant to solve it. As expected, the response appears independent of ratio.

An ideal, noise-free perception of magnitude feeding an errorless decision-making system would result in magnitude discriminability coefficient w equal to zero. However, here we observed relatively large Weber fractions in the duration- and distance-discrimination tasks (Figure 1b) suggesting that in the monkey brain these processes are quite noisy. The moderate error rate in the match-to-sample task, in which magnitude perception is not essential, indicates that other noisy mental processes (e.g., memory storage and implicit task selection) contribute to the variability of response-selection in this and the other task(s). To account for the variability of these non-perceptual processes, we subtracted a noise term $\sigma_{non-perceptual} = 0.15$ from the calculated behavioral discriminability coefficients w (shown in Figure 1b). This permitted us to estimate the variability coefficients of the internal representations of duration-discrimination, w_0 = 0.34 and distance-discrimination, w_0 = 0.33. We used these parameters to generate noisy magnitude stimuli using the aforementioned *number-line* model $\hat{\eta} = N(n, nw_0)$ - but note that two control simulations reported below show that the learning mechanism is robust to greater levels of noise. Finally, we added constant variability at response-selection by randomly alternating 5% of the responses.

Information-theory measures used in the analyses To compare the PFC neurons in the (Genovesio et al., 2012) study and the clusters evolved by the computational model we used information-theoretic measures. In particular, we analyzed the *information content* (in terms of properties of interests such as goals or colors and their combination) conveyed by the PFC neurons and the clusters evolved by the computational model.

In order to apply information measurements to responses with many levels (or continuous values), the responses first need to be discretized at just few levels (e.g., two or three); this procedure is necessary to avoid that sparse observations of multiple response levels artificially distort the information measures (Panzeri et al., 2007). We adopted a simple three-level response-discretization procedure based on normalized values: first, we normalized the response by subtracting its mean and dividing by its variance; then, we created three categories separated by levels -0.5 and 0.5. A preliminary entropy-measuring analysis revealed that the three-level discretization increased the overall information content relative to a simpler two-level discretization; furthermore, it did not lose too much information relative to a four-level discretization.

We used the three-level discretization to perform information-criteria analyses of two kinds of raw responses: the firing rate of the PFC neurons and the selection probabilities of the latent states of the computational model. The properties of interest had already just few levels: the goal had two levels (red and blue target); the task had three levels (*duration*, *distance*, and *match-to-sample*); the index of the larger stimulus had two levels (either S1 or S2), and the color of the first stimulus also had two levels (either red or blue).

The information-theoretic measures we used are introduced below:

Entropy $H(x) = -\sum_i p(x_i) \log p(x_i)$ is a measure of the overall quantity of information conveyed by response x and it essentially measures response variability. Here, $p(x_i)$ is the probability of each specific response level x_i.

Mutual information $I(x;s) = \sum_{i,j} p(x,s) \log_2 p(x,s)/p(x)p(s)$ measures the information carried by the neural response x about a stimulus property s where $p(s)$, $p(x)$, and $p(x,s)$ are the marginal and joint empirical probabilities of the property and the response. Critically, in the case of multiple related properties, it does not measure the specific amount of information carried by each of them. In our task, this might potentially confound the interpretation of goal-coding neurons. Because goals could possibly encode a mixture of color, magnitude, and task information, it is possible that neurons encoding one of these properties carry non-specific information about the goal.

To rule out such confounds, we used *conditional mutual information* $I(x;s|s') = \sum_k p(s') \sum_{i,j} p(x,s|s') \log_2 p(x,s|s')/p(x|s')p(s|s')$ that measures the amount of information about a property s while controlling for another property s'. Note that relative to $I(x;s)$, $I(x;s|s')$ can decrease, remain invariant, or increase. The more it decreases, the more the response encodes s by virtue of s'. Following our goal-coding hypothesis, we expected that the mutual information between the response x and the *goal* property conditioned on all related properties would not drop to (or be close to) zero. Indeed, such a drop which would imply that most of the goal-related information is explained away by the property s'; for example, if $I(x;goal|colour)$ drops close to zero, then color-coding would be a more parsimonious explanation than goal-coding. To verify that the conditional information is statistically different than zero, we used the formal non-parametric method of Ince et al. (2012).

Tuning functions and Contrastive Preference of neurons for stimuli features

We finally considered the PFC neurons and the clusters evolved by the computational model as

neural filters and analyzed the preferred stimuli features to which they may had been tuned. The *preference* p_i^f of a latent category z_i for a given feature f was indexed with the expected conditional probability that the set of context stimuli having such feature $c_f = (\forall\, c: f(c))$ would activate the category z_i:

$$p_i^f = {}^1/_{|c_f|} \sum_c P(z_i|c) \qquad (4)$$

where $|\, c_f \,|$ is the number of stimuli having feature *f*. The index ranges from 0 to 1, with value of 0 meaning null preference and value of 1 indicating maximal preference.

As a second step, as in our testbed monkey experiment (Genovesio et al., 2012), we expressed the preference to one of a pair of related features (f_1, f_2) by calculating the *contrastive preference* $p_i^{f_1-f_2}$ - the difference between the preferences for each of those features, see equation (5). The index ranges from -1 to +1, with value of +1 meaning preference to f_1, value of -1 meaning preference to f_2, and value of 0 meaning no specific preference.

$$p_i^{f_1-f_2} = p_i^{f_1} - p_i^{f_2} \qquad (5)$$

RESULTS

The probabilistic computational model was first trained and then tested in a block design following the target monkey experiment protocol. Each block presented pseudo-randomly selected patterns from a given task, with noise added as described earlier.

Below we report the results of the simulations of the monkey experiments (behavioral and neural data) and of several control experiments. All the results reported below are calculated as the average of 30 simulations.

Behavioral results

The first critical test for our model was the ability to adequately replicate the monkey behavioral data. Using the monkey protocol and applying perceptual and action-selection variability measures derived from the analysis of the monkey psychometric profile (see the Methods section), we administered the three experimental tasks to 30 replicas of the model, which were learned successfully within a few thousands of trials (Figure 7) despite the considerable noise of the stimulus-action-value contingencies. Response accuracy during the test period was equivalent to monkeys' performance: 80% (s.e. = 0.5%) in the duration task, 80% (s.e. = 0.5%) in the distance-task, and 95% (s.e. = 0.4%) in the match-to-sample task (reward-driven learning continued also in this test period). At the behavioral level the model exhibited monkey-like response accuracy, psychometric profiles (per-task correlations, R^2>0.88) and magnitude discriminability in the test period (Figure 1b,d).

Neural-level analysis

Following an adequate simulation of monkeys' behavior (Figure 1b,d) we performed a neural-level analysis of the way the model obtains flexible control of behavior. In keeping with the goal-coding hypothesis, we predicted that 1) the model would have learned the observed stimuli-action-

value contingencies by clustering the large number of context stimuli into a small set of latent states or categories; and 2) that the latent states would correspond to goals (i.e., the color of the target stimulus that had to be selected). The other possible clustering structures were a common representation of magnitude (e.g., clusters coding for "the larger" and "the smaller" stimuli) or a sensory oriented encoding, clustering one or a combination of context stimuli properties (e.g., cluster coding for colors).

We found that the computational model consistently used a very limited number of popular categories, each aggregating a large number of contexts (Figure 2a,b). However, this result per se is not yet sufficient to assess the goal-coding hypothesis – that is, that the used latent categories actually carry goal-related information; to this aim, it is essential to analyze what these latent states encode. The next relevant questions were: are the learned categories purely perceptually driven (Freedman et al., 2001)? Did they code basic perceptual categories such as stimuli magnitude? Or did they cluster prospective goals or control signals? Did they correspond to the coding characteristics of the monkey PFC neurons studied in the same tasks? We approached these critical questions with a thorough information-criteria analysis based on (conditional) mutual information measures.

Analyses using information measures

We used information-theoretic measures to assess the coding properties of the PFC neurons in the (Genovesio et al., 2012) study and the clusters created by the model, and in particular to assess if they code (or carry information on) prospective goals, in keeping with the goal-coding hypothesis. Here, "goal information" indicates a property or a set of properties that are relevant for a (future) choice, e.g., whether red or blue should be the choice for the target. In the monkey experiment, there are two possible targets for the choice, so goal information can be measured as a contrastive preference (see formula 5) between the to-be-chosen vs. the not-to-be-chosen target (e.g., red vs. blue).

We first analyzed all the 324 PFC-neurons recorded from the three tasks and all latent categories whose activity conveyed at least 0.8 bits overall information (measured with entropy) and 0.10 bits mutual-information about the goal (arbitrary thresholds). This analysis identified n=117 latent categories (among the 30 replicas of the model) and n=31 PFC neurons encoding goals (Figure 2c,d; black dots indexing the mutual information conveyed by the unit response about the goal).

However, as explained in the methods section, the mutual information analysis cannot rule out the possibility that goal-coding is "spurious" and confounded by other properties of the stimuli properties. To verify whether these units conveyed genuine goal information, we then calculated the conditional mutual information conveyed by the units about the goals, considering various task-related potential confounds of the goal-property (Figure 2c,d; color dots indexing the corresponding conditional mutual information for each unit). As explained in the methods, if the mutual information conditioned on a given property substantially decreases (relative to the non-conditional mutual information) and approaches zero, then this property would *explain away* the information conveyed by the unit about the goals (or in other words, the units would encode a confounding property, not a goal). However, for all the units identified as goal-coding, the conditional mutual information was significantly greater than zero ($p<0.05$ using the method of Ince et al., 2012), revealing thus genuine domain-general representation of prospective goals in both the PFC neurons and the latent categories, although noisier in the neural data.

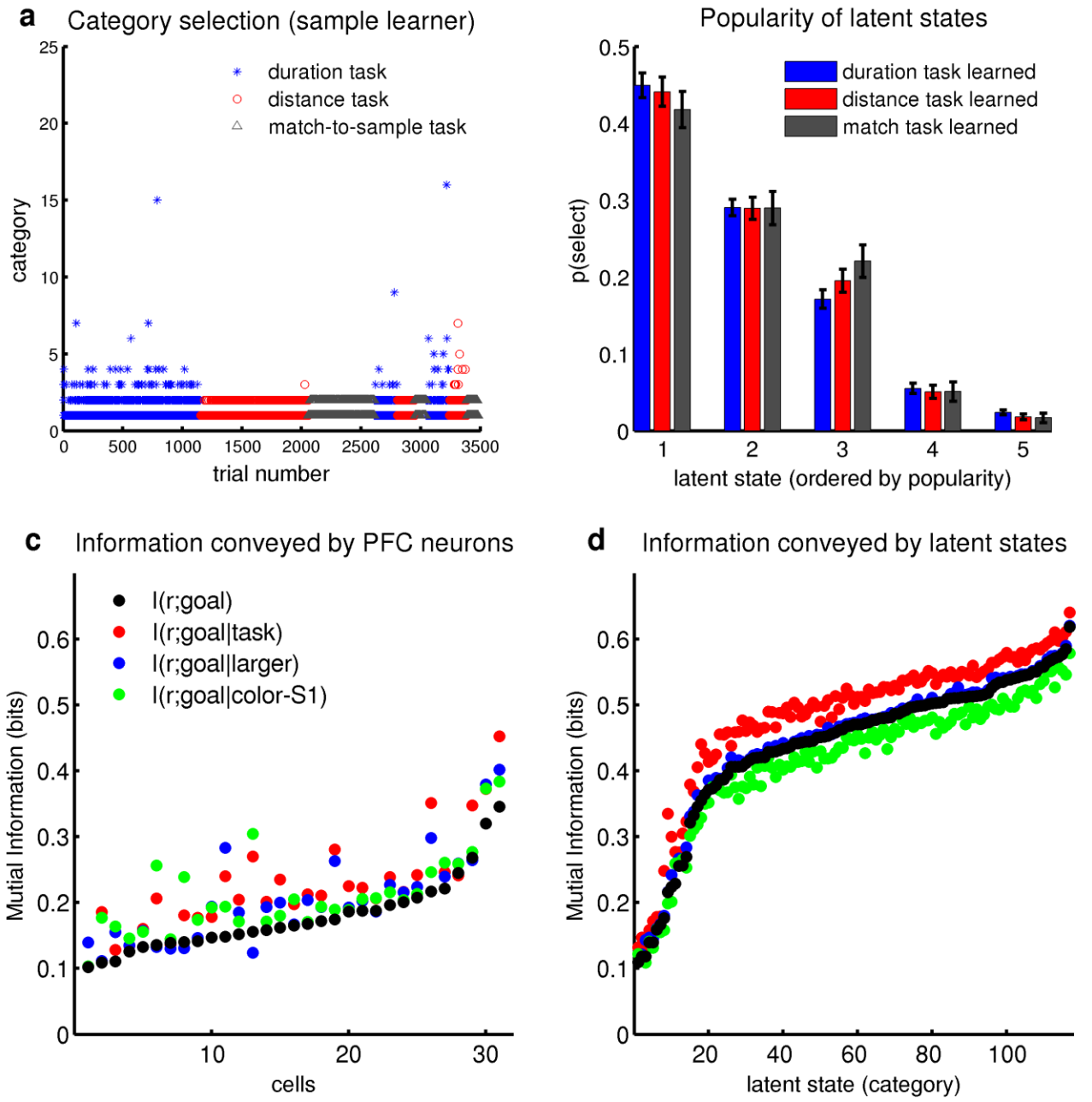

Figure 2: Neural-level analysis. (**a-b**) Learning evolved compact latent-state control scheme favoring transfer learning: (**a**) raster plot indexing the selected context-category during each trial. (**b**) Popularity of the most frequently used latent states (bars, s.e.). Note that all the three tasks use the same (few) latent states. Once a role for a given latent state is established in the first task, it remains unvaried in the successively learned tasks – a mechanism that might play an important role for transfer learning, see Figure 7. (**c-d**) Information-theoretic analysis of PFC-neurons (**c**) and latent categories (combining the latent units of all 30 replicas) (**d**) showing the amount of overall information conveyed by the selected units about the goals (black dots; units ordered by the amount of mutual information) and the amount of goal-information when conditioned on various task-related properties (color dots). Since for each unit the conditional mutual information does not (or just slightly) decrease relative to the mutual information, both the PFC and the model genuinely encode goals independent of basic task-related properties; the model does so less noisily, probably because of lack of previous experiences. In both the PFC data and the model, I(r;goal | task) is slightly higher than I(r;goal), see Figure 2c,d. In the model, I(r;goal | color S1) is slightly lower than I(r;goal). To understand why this is the case, it is necessary to note that Figure 2d shows the average of the three tasks, and in the match-to-sample task I(r;goal | color S1) is close to zero, because the identity of S1 is sufficient to infer the goal, or in other words S1 explains away the goal information. Also in the PFC data, I(r;goal | color S1) is close to zero; the reason why the average value shown in Figure 2c is slightly higher is because of the slightly higher (and noisier) value in the other two tasks.

Tuning functions

To corroborate this conclusion, we investigated the coding properties of the model units and PFC goal-coding neurons (as identified using the above analysis), by compactly expressing goal-coding for each task in terms of *contrastive preference* (see formula 5) to respond to the color corresponding to the prospective goal (e.g., "red") but not the other (e.g., "blue"). Figure 3 shows the scatter-plots of preferences, calculated separately for each pair of tasks. Note that the dots representing task-sets are exclusively present in the top-right and bottom-left panels and lay along the main diagonal. This result indicates that, like in the monkey data, the goal-coding responses are task-invariant (i.e., the task sets have the same goal preference in all the three tasks), which corroborates the hypothesis of a domain-generality of goal-codes. Note that the coding properties of goal-codes are different from (and cannot be explained in terms of) simpler color-selective – or color-coding – responses, because the neurons/units were only active when the color they were selective for corresponded to the behavioral goal (see Figure 4). This result corroborates the hypothesis that latent categories are true goal-codes rather than having a mere preference for sensory properties like color or magnitude.

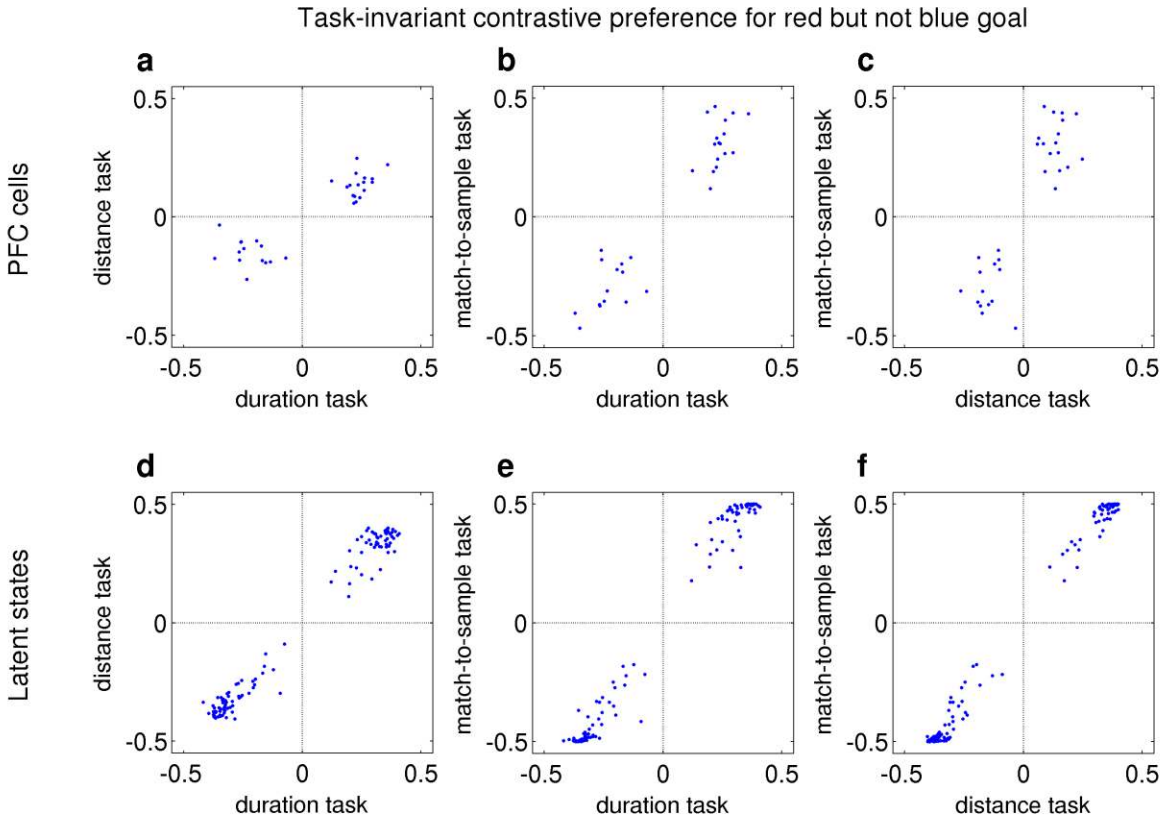

Figure 3: Goal-coding: PFC neurons (**a-c**) and latent categories (**d-f**) exhibit the same task-invariant goal preference. Each point indexes the normalized contrastive preference of a single unit to one but not the other goal (aggregated data). (See also Figure 4)

In a further analysis we also investigated the coding properties of the 20 most frequently selected clusters in every replica of the model, and in particular if they encoded one or more of the following properties: *goal-coding* (Figure 4a,b); *order-based magnitude preference*, by calculating the contrastive preference for greater first (S1>S2) and second (S1<S2) stimuli magnitude (Figure

4d); and *color preference* of one of the context stimuli, S1, by calculating the contrastive preference for "red" and "blue" S1-stimuli (Figure 4g).

Once again we found that the goal-coding preference was task-invariant (Figure 4a,b), providing thus strong evidence about its domain-generality. Moreover, and consistent with monkey data, no latent category units exclusively encoded other stimuli dimensions such as a common representation of magnitude across tasks (Figure 4d).

Overall, the emerged goal-codes were not purely stimulus-related but constituted a task-relevant abstraction: a possible way the PFC might convert order- and feature-based stimulus codes into a domain-general but goal-specific code that affords efficient action selection.

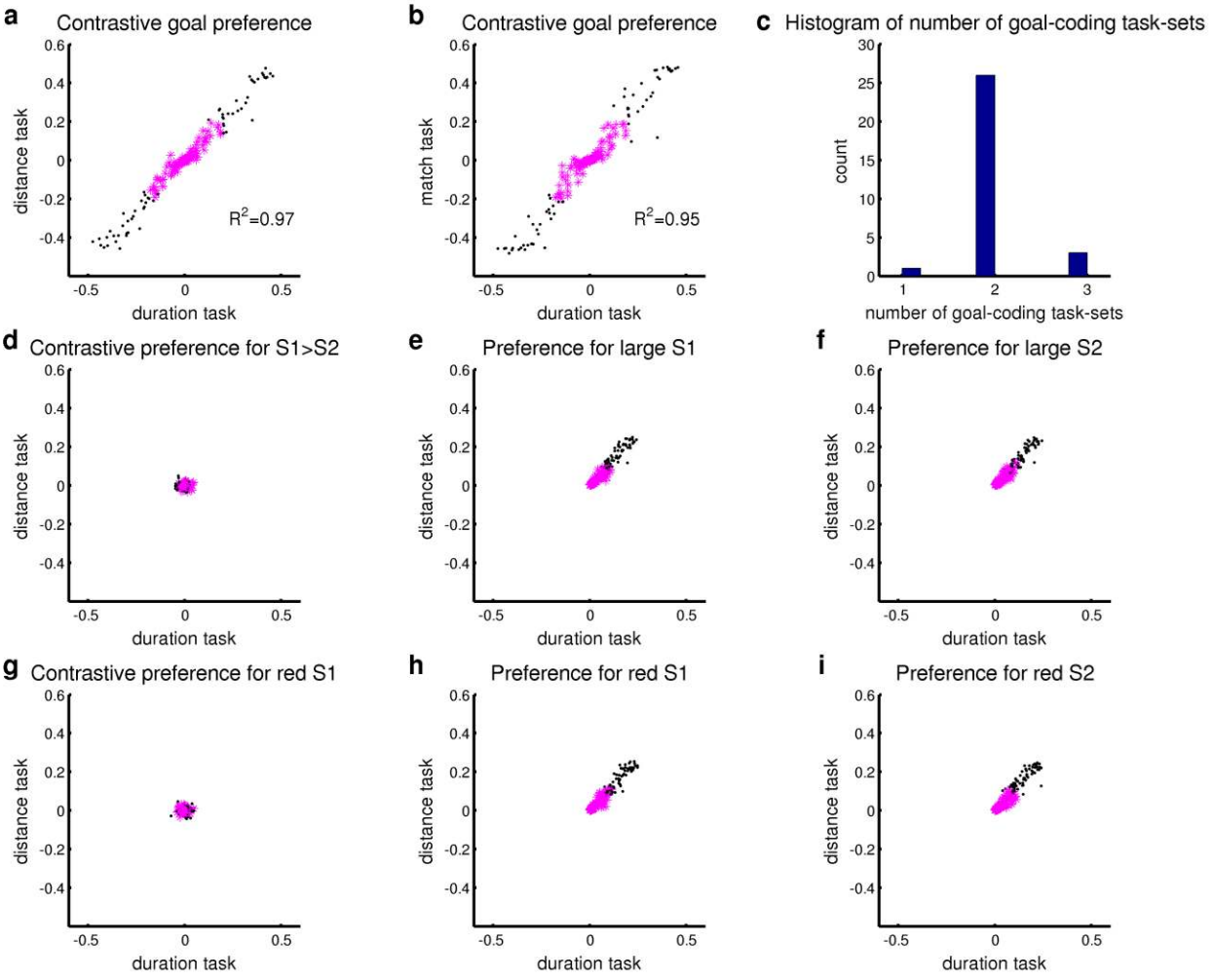

Figure 4. *Preferences* of the 20 most frequently selected latent categories for each of the 30 learners (i.e., without information-criteria selection). Each point in panels (**a,b,d,g**) indexes the non-normalized contrastive preference of a given latent category in two different tasks (axis labels) for various combinations of features (Methods). Values at the top-right and bottom-left angles indicate the same preference, and values near the center indicate lack of preference. The most clear is the goal-preference (panels **a,b**) in corroboration of **Figure 3,d-f**. Black dots and magenta stars indicate strong and weak goal-preference (arbitrary threshold of 0.2), respectively. Consistent with monkey data, no latent category exhibited preference for order-based relative magnitude coding (**d**) or for color (**g**). Instead, the categories with strong goal-preference (black dots) also show a weak preference for simple properties (magnitude-size and color, **e-f,h-i**) that combine to build goals. A histogram of the number of latent-categories with strong goal-preference per learner is shown in (**c**).

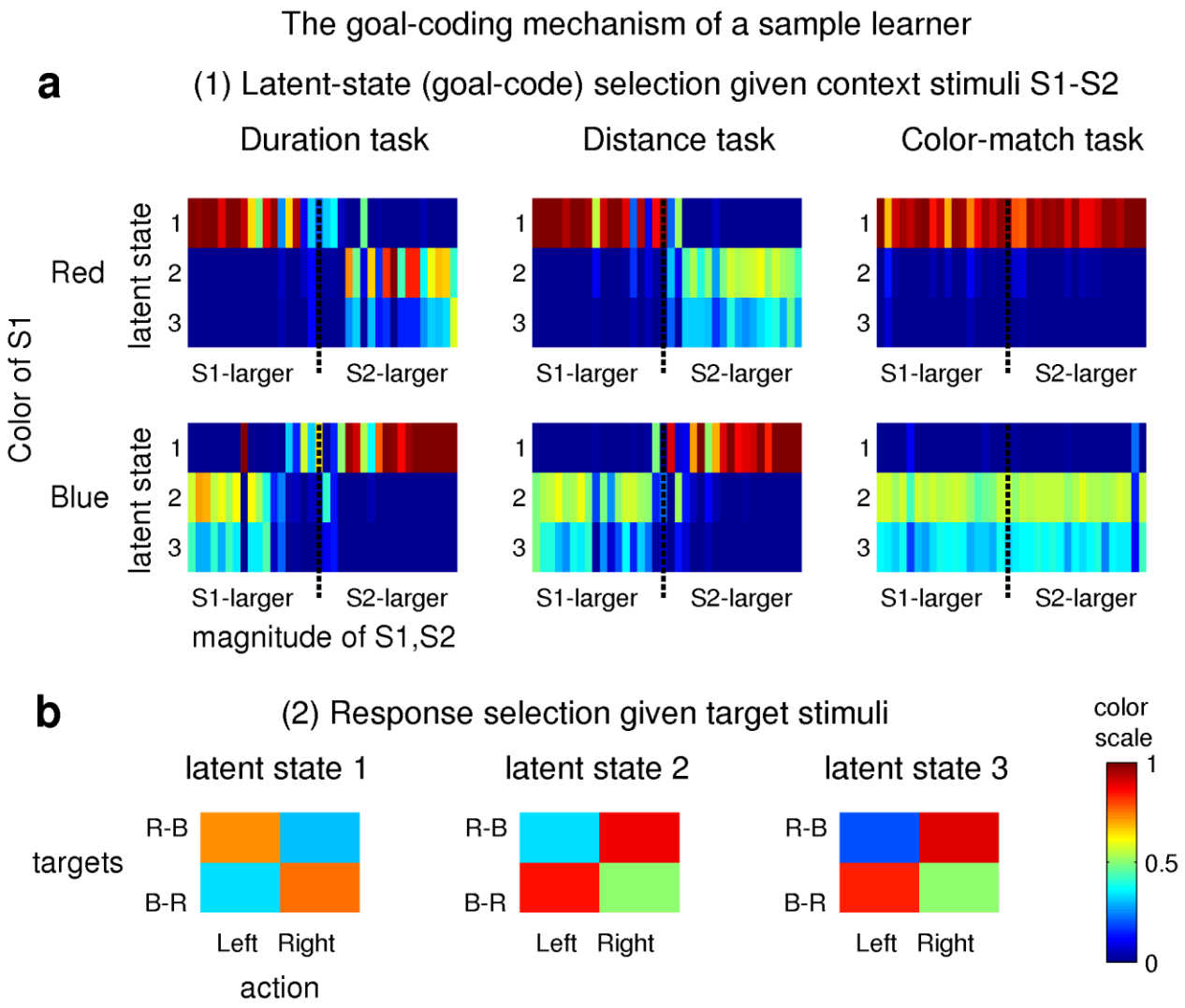

Figure 5. Emergent goal-coding mechanism in a sample learner. (**a**) In each of the three tasks studied here (duration, distance, or color-match), context stimuli S1 and S2, each characterized by a color (either red or blue) and variable magnitude properties (either duration or distance) are clustered into few popular latent states (here, the first three are shown) that correspond to the color-identity of the to-be-selected target stimulus (red or blue). The six colored matrices show latent-state selection in each of the tree tasks and for each color condition (S1-red or S1-blue; the color of S2 is either complimentary to S1 in the Duration and Distance task, or equal to that of S1 in the Color-Match task). Each colored matrix shows the probability (blue-to-red gradient indexes probability from 0 to 1) of selecting each of the three most-popular latent states (row 1, 2, and 3) for each combination of relevant magnitudes characterizing S1 and S2 (columns; ordered by decreasing difference between the magnitudes of S1 and S2). In this learner, the first latent state is preferred when the greater stimulus is red or in case of red color-matching stimuli ("red goal"). The second latent state is preferred when the greater stimulus is blue or in case of blue color-matching stimuli ("blue goal"). Note that the second and third latent states are almost equally preferred in the same conditions, which implies that the blue-goal was redundantly encoded. This redundancy is controlled in the model by the concentration parameter alpha (see Method section). (**b**) Given a selected latent-state, say red-goal, the response selection depends on the location of the corresponding target (i.e., whether the red target is to the left or right). In other words, if latent state 1 has been selected (i.e., red goal), a red target located to the left (or right) evokes a left (or right) target response. If instead latent states 2 or 3 have been selected (i.e., blue goal), left- and right-located blue targets correspondingly evoke left/right switch presses.

Analysis of the emergent goal-coding mechanism

To better understand the principle of goal-coding, we analyzed and showed in Figure 5 the emergent neural-level mechanism of goal-coding and response selection in one sample learner, whose latent-state selection is shown of Figure 2a. At the top is shown the probability $P(z_i|c_j)$ of selecting each of the three most-popular latent states z_i for each combination of context-stimuli c_j properties (color and duration/distance of stimuli S1 and S2), separated in six different panels by the implicit stimuli-dependent task and color of S1. The bottom of the figure shows the probability $P(r=1|s_k,a_l,z_i)$ of obtaining reward given each combination of target stimuli s_k (Red-Blue or Blue-Red) and possible action a_l (press Left or press Right), separately for each of the three most-popular latent states z_i whose selection preference is shown above. This probability is used by the learner to select the action that brings reward.

Thus, upon the (noisy) perception of context stimuli S1 and S2, this learner would preferentially select the first latent state if the stimuli were perceived as having red goal (i.e., larger or longer red stimulus or red color-matching stimuli), and the second or the third latent state if a blue goal was instead perceived (i.e., larger or longer blue stimulus or blue color-matching stimuli). The selected latent state conditions the successive response selection upon target appearance. For example, if the red goal was selected (i.e., the first latent state is active), and the target stimuli correspond to "red to the left, blue to the right", the learner would press the left button in order to obtain reward.

Control simulations

The simulations thus far replicated the monkey data in the specific testbed conditions at both the behavioral and neural levels. In addition, we have conducted a set of control simulations, with two aims: 1) demonstrating the generality, robustness and scalability of our computational methods; and 2) generating novel empirical predictions – a paramount feature of computational modeling, whose results should extend beyond the mere replication of existing data.

Effects of perceptual noise. An important question is whether the proposed computational scheme is robust to various levels of noise in magnitude perception, or in other words whether it permits extracting appropriate goals with high or low levels of noise. We investigated this issue in two control simulations. In one, the noise was half of that used in the main simulation, while in the other the noise was doubled. In general, we expected to observe corresponding changes in behavioral discriminability, but not qualitative differences such as the impossibility for the model to extract goal-codes (unless of course the model experiences ceiling effects of noise, which would preclude learning in general, not only goal-coding).

As predicted, much smaller perceptual variability (w_0=0.17 for both distance and duration) largely improved magnitude discriminability at the behavioral level (duration: w=0.28, s.e. = 0.01, distance: w=0.27, s.e. = 0.01); and neural-level analysis revealed the same control mechanism based on goal-coding (Figure 6a-c). Note that this simulation would represent a closer approximation of an experiment with adult human participants, whose magnitude (or analog number) processing system is generally more precise than that of monkeys.

The result of the second control simulation (with higher levels of noise) exceeded our expectations. The subjects learned to identify and select correct targets despite very large perceptual noise (: w=0.68 for both distance and duration) and consequently, with much worse magnitude

discriminability relative to the main simulation (duration: : w=0.99, st. err = 0.039, distance: : w=0.89, st. err = 0.034). Importantly, a neural-level analysis revealed that even in this case the model extracted goals during the first two, noisy tasks (Figure 6d-f), thus demonstrating the robustness of the goal-coding principle.

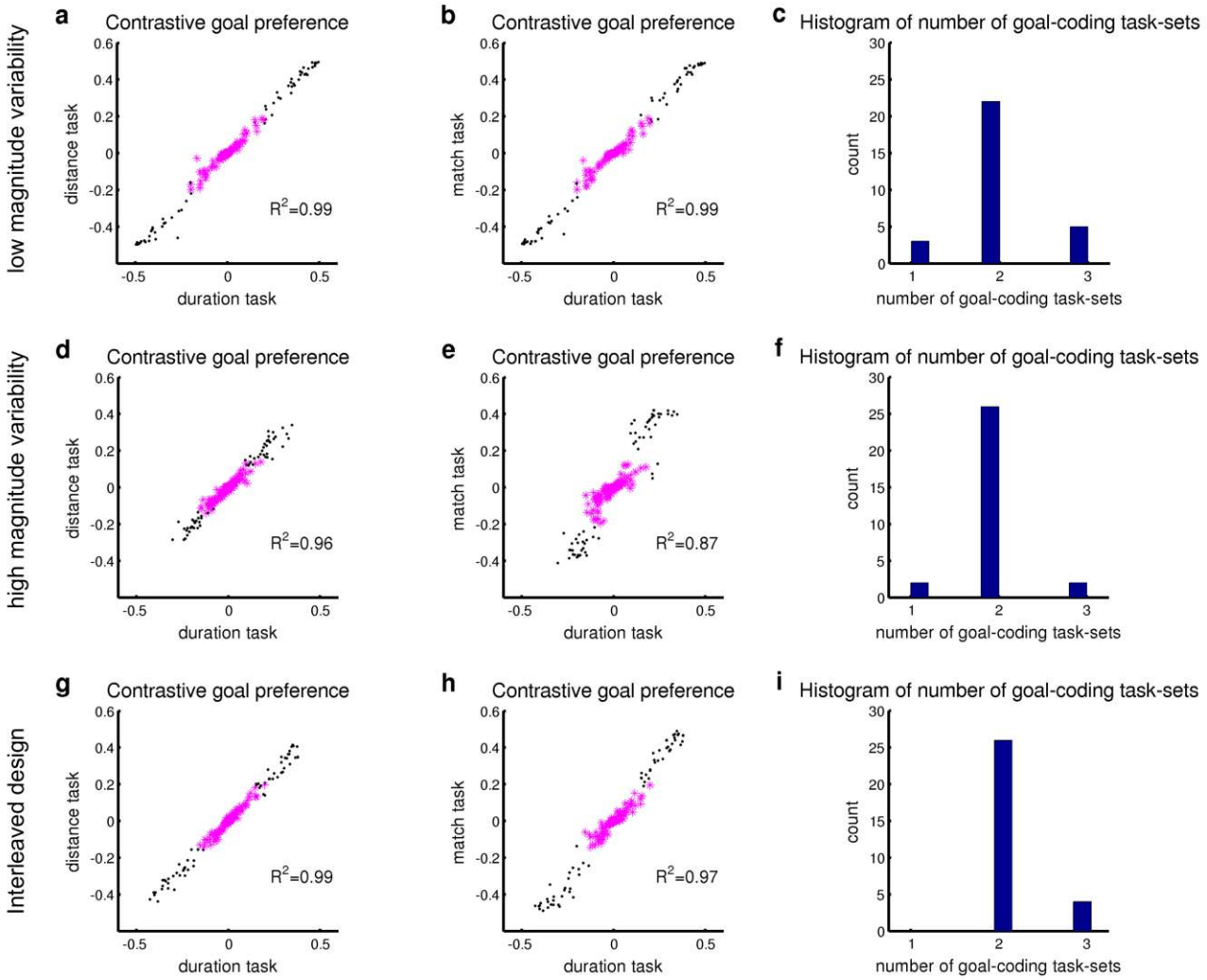

Figure 6. *Robustness* of the goal-extraction mechanism for cognitive control revealed by control simulations in various easier and more challenging learning conditions. (**a-c**) Small variability of magnitude perception; block-design learning. (**d-f**) Large variability of magnitude perception; block-design. (**g-i**) Variability as in the main simulation but tasks learned and tested in completely interleaved design. Each point in panels (**a-b,d-e,g-h**) indexes the non-normalized contrastive goal-preference of a given latent category in two different tasks (axis labels). See also Figure 4 and Methods.

Redundancy of goal-coding. We then verified the role of the concentration parameter α that controls the probability for a newly experienced context to evoke a new latent state ("table" in the Chinese Restaurant) or to join some of the popular latent states. Control simulations with various levels of this parameter (α = 1, 2, 5, 10) revealed that the performance, sensitively measured with the behavioral Weber fraction w, was essentially unaffected by α. At the same time, the number of exploited latent states increased along with α, as expected. However, for a given level of alpha, the same (popular) latent states were used in all three tasks, which together with contrastive-preference analysis of these latent states as that in Figure 3 corroborated the finding in the main simulation that once a specific

(goal-coding) role of a latent state is established, it remains invariant in the successfully learned tasks, which in turn is critical for transfer learning. Finally, with the help of our information-theory analysis we also found that the goal-coding specificity was essentially the same as in the main simulation. Thus, the concentration parameter alpha essentially controls the redundancy of goal-representation, but it does not change the goal-coding principle.

Goal-coding selection criterion. We further verified the impact of the specific selection criteria on our novel neural-level analysis. Putting a threshold on entropy was necessary to ensure that the analysis focused only on units having a reasonable amount of response variability. Halving the threshold, from 0.80 to 0.40, left unchanged the pool of goal-coding PFC neurons and extended the pool of latent states by just five units, obtaining essentially invariant results. We expected a more significant impact of the mutual-information selection criteria that specifically sought goal-coding units. Indeed, doubling it, from 0.10 to 0.20, restricted the selected pool to just n=9 PFC neurons and slightly decreased the number of latent states. Expectedly, these very specific goal-coding units showed more clearly dichotomous contrastive goal-preference distribution. On the contrary, halving the mutual-information criterion, from 0.10 to 0.05, extended the pool to n=53 PFC neurons and slightly increased the number of latent states. As expected, this less-specific pool had slightly less-clear dichotomous distribution of contrastive goal-preferences.

Effects of block design vs. interleaved design. The main simulation presented the stimuli in block-design, but we hypothesized that the same goal-based control mechanism would emerge without task blocking, as well. To test this prediction, we conducted a control simulation in which the tasks were learned using *interleaved design* by pseudo-randomly selecting task-type in each trial; all the rest was kept invariant. The result confirmed the prediction. Relative to the main simulation, magnitude discriminability just slightly increased (duration: : w=0.54, st. err = 0.015, distance: : w=0.50, st. err = 0.012), but goal-coding and the associated control strategy was consistently found (Figure 6g-i).

Transfer learning and the differences between standard (flat) RL learning and our proposed (structured) method. Classical reinforcement learning methods (e.g. TD learning) are sufficient to learn the correct policy in our tasks, provided that each pair of context-target stimuli is observed several times. However, we hypothesized that our proposed (structured) method based on a non-parametric component - which essentially extracts goal-to-response mappings - would have been advantageous when learning novel tasks that share similarities with the already acquired ones (i.e., *transfer learning*) – pointing thus to a specific adaptive value of structured models and goal-coding in the PFC.

To verify this prediction, we run a control simulation using a flat generative probabilistic model of reward $P(r_t \mid s_t , a_t , c_t)$, with the same binomial distributions as in the main model. As hypothesized, the flat model successfully learned the tasks. However, the flat model showed a slow learning process that has the same trend for each new task, with poor or no generalization. This is in contrast with the learning trend of the structured model that instead reuses its knowledge to learn faster each novel task (Figure 7). Thus, a key advantage of the non-parametric component is the predisposition to build and reuse already acquired goal contingencies across different domains and in novel situations, which is consistent with a role of PFC in supporting one-shot learning and providing behavioral flexibility without catastrophic forgetting (Doya, 1999; Koechlin and Hyafil, 2007; Shima et al., 2007).

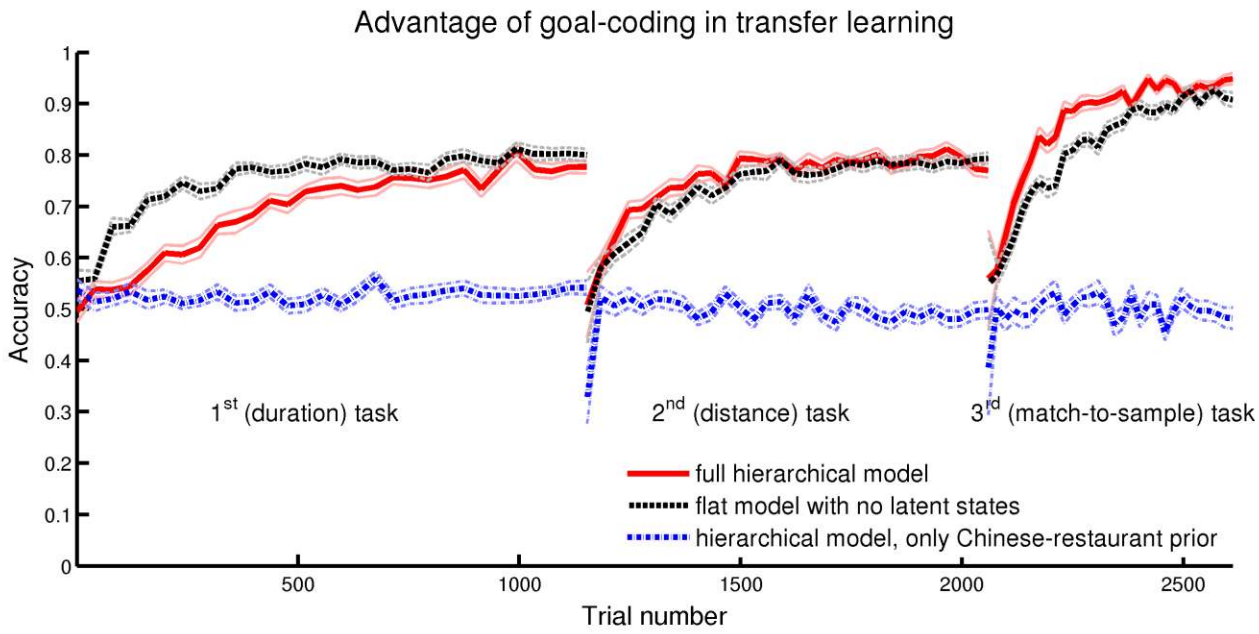

Figure 7. *Advantage of goal-coding in transfer learning*. Comparison between: (i) the learning trend of the non-parametric (full) hierarchical model in which goal-codes emerge (i.e., the main simulation; thick red line), (ii) the learning trend of a control *flat model* in which the learning task for every context-target pair departs from the same uniform prior (thick black dotted line), and (iii) the learning trend of a control (hierarchical) model with *untrained latent states* (i.e., using only the Chinese-restaurant prior; thick blue dashed line). Relative to the flat model, the initial delay in the learning trend of the full hierarchical model (due to initial learning of latent categories) turns into a consistent advantage during the learning of successive tasks: a signature of transfer learning. The control model was not able to learn the tasks, plausibly because it lacks a mechanism that extracts useful internal representations in a restricted feature set. Thin lines delineate s.e. bands.

Unsupervised clustering and the role of supervised category learning. To show the critical role of categorization driven by the reinforcement signal, we performed another control simulation in which the latent units were initialized according to the Dirichlet mixture process (as in the hierarchical model used in the main simulation) but not further trained to account for the conditions bringing to reward (i.e., not applying formula 2). The procedure, parameters, number of replicas, learning schedule were the same as in the main simulation. We expected that the Chinese-restaurant-process prior would randomly associate the input context stimuli with a limited number of active latent variables, providing no useful internal representation of the context and thus producing low performance. Indeed, as shown on Figure 7, this model was not able to learn the tasks and responded at chance level, further emphasizing the importance of value-driven learning in shaping a behaviorally relevant categorization process.

Scalability of the computational learning approach. To assess the scalability of the proposed method to more challenging experimental conditions, we generalized the setup by increasing the number of colors and available responses. The identity of each context stimulus S1 and S2 were randomly drawn among k colors, and the target stimulus was a random permutation of all available identities (colors). As in the main simulation, the action consisted in selecting the identity of the stimulus that (as in the main simulations) lasted longer, was more distant, or matched the context stimuli, but this time located in one of k possible positions and among all other identities. Note that

the complexity of the task increases dramatically along with ***k***. For ***k***=3, there are 648 context stimuli and 6 target displays, while for k=4, there are 1152 context stimuli and 24 target displays.

The model successfully learned also these two problems, after a higher number of learning trials reflecting the increasing complexity (k=3: 20.000 trials, k=4: 50.000 trials). As in the main simulation, the reinforcement learning procedure used a goal-coding strategy to solve the tasks and the goal-response mappings created in the first phase greatly simplified the learning of the second and the third tasks (Figure 8), further supporting the generality of the proposed approach.

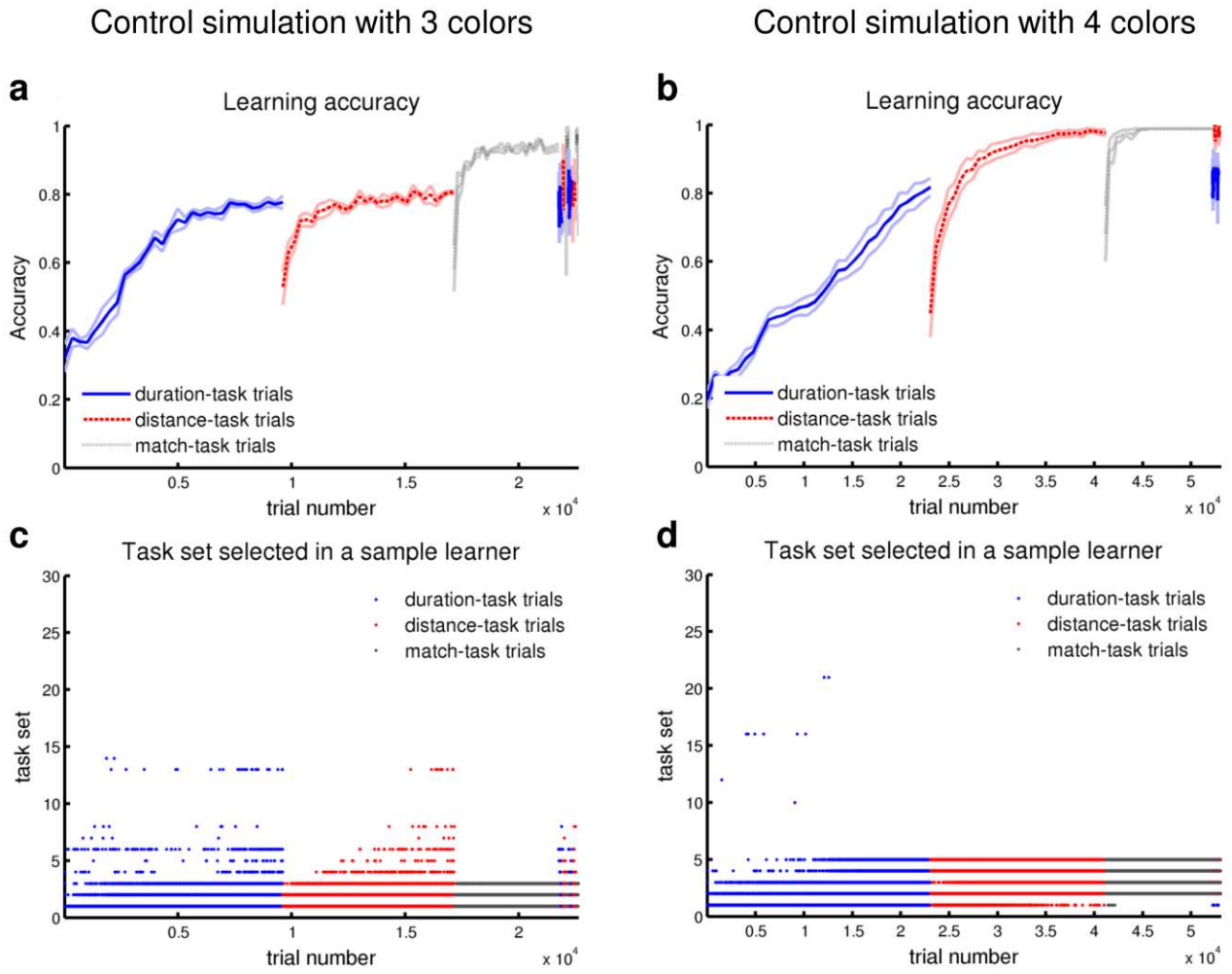

Figure 8. *Generalization:* learning trend (**a,b**) and raster plot of category selection (**c,d**) of control simulations generalizing the tasks to more than two stimuli identities (i.e., colors; **a,c**: k=3; **b,d**: k=4), demonstrating scalability of the method. High learning accuracy is obtained at the cost of increased number of trials. The number of used categories increases relative to the main simulation (**Figure 2a**), consistent with the increased number of possible goals.

DISCUSSION

The PFC lies at the apex of the brain control hierarchy (Fuster, 1997) and is uniquely positioned to integrate context, reward, and control-related information and to learn their (noisy) contingencies. This gives PFC great flexibility in supporting goal-directed behavior but also implies that it has to solve complex, multidimensional learning and selection processes. From a statistical viewpoint, this

problem can be finessed by learning a hierarchical generative model that links stimuli, actions and rewards to internal ("hidden" or "latent") states or categories, which need to be inferred, too (Friston, 2010). In keeping with this idea, in tasks requiring subjects to learn a large set of actions the PFC was found to develop abstractions, or categories of actions that permit to guide the behavior (Shima et al., 2007); however, other kinds of categories have been found in the PFC such as object categories (Freedman et al., 2001), leaving open the question of what kind of categories better support goal-directed behavior. Here we tested the idea that goal-codes or prospective representations of goals constitute a solution to the problem faced by the PFC: learning abstract categories that are useful to steer goal-directed action and cognitive control. We hypothesized that goal representations (or goal-codes) – here, prospective representations of the to-be-selected target stimuli - emerge in PFC as "latent states" (or categories) of a generative model that clusters relevant statistical properties of stimuli and value information and successively bias response selection towards goal-relevant outcomes.

To test the hypothesis we used a probabilistic generative model that combines unsupervised non-parametric learning (for latent state learning and categorization) and reinforcement learning (to guide the categorization towards task-relevant abstractions). Non-parametric Bayesian networks have flexible structure allowing learning rich internal representations of complex data (Ghahramani, 2013). Previously, approximate non-parametric learning successfully developed categories (Sanborn et al., 2010), and non-parametric value-driven learning was used to build task-sets (Collins and Koechlin, 2012; Collins and Frank, 2013) supporting the viability of the method.

The testbed for our simulation was a series of studies reported in Genovesio et al. (2012), where the experimenters collected monkeys dorsolateral and caudal PFC single-cell data during the post-learning period, and reported goal-coding cells common to all three tasks. The results of our computational simulations and information analyses successfully replicated these data; and beyond that, they showed that goals emerge as latent task dimensions that encode behaviorally relevant task regularities and stimuli properties, thus offering a normative explanation for the domain-general representation of prospective goals found in the monkey PFC (Genovesio et al., 2012; Yamagata et al., 2012).

Furthermore, despite the Chinese restaurant process (CRP) produces a very high number of latent states and usually selects (or "populates") a logarithmic function of the number of experienced context stimuli (Gershman and Blei, 2012), we found that our model – which uses reinforcement signals in combination with CRP – consistently uses very few of them. This result is consistent with the idea that, while sets of broadly tuned PFC neurons might provide a "basis" or "repertoire" to execute a variety of cognitive control tasks, each specific task might critically depend on a smaller set of cells that have highly selective and task-specific (e.g., goal-related) properties. Future studies looking at the dynamics of PFC representations during learning might permit testing this hypothesis and studying if the learning process benefits from the putatively "critical" properties to be already present in a PFC "repertoire" (transfer learning) and/or from *adaptive coding:* the ability of PFC neurons to flexibly adapt their properties to convey task-relevant information (Duncan, 2001).

The close matching of the model and the data at both behavioral and neural levels, and the results of our control simulations in more challenging experimental conditions, support our hypothesis that PFC goal-coding might be a fundamental organizing principle for efficient flexible control. Furthermore, our novel analyses based on information-theory measures corroborate the goal-coding hypothesis by ruling out the possibility that the neuronal coding of goals was the result of a confound with other task-related features.

Key to our results is the combination of two forms of learning in which an unsupervised category learning process extracts relevant behavioral abstractions from the input data but the selected categories are sculpted by a value-driven process according to their adaptive value (see Equation 2). Both forms of learning have been extensively reported in the PFC (Frank and Badre, 2012) but they are typically studied in isolation. Our proposal thus brings an integrative perspective that reconciles two influential streams of research on prefrontal function that focus on behavioral control (Miller and Cohen, 2001; Passingham and Wise, 2012) and category learning (Seger and Miller, 2010), respectively.

Furthermore, at difference with most (model-free) reinforcement learning models of cognitive control that use direct stimulus-response mappings and in which *goals* are implicitly encoded in a value function of states and actions (Botvinick et al., 2009; Sutton and Barto, 1998; Dayan, 2009; O'Reilly et al., 2010), in our method goals are explicitly coded. Explicit goal representations are a characteristic feature of most model-based probabilistic architectures for goal-directed behavior, such as *planning-as-inference* (Pezzulo, 2012; Solway and Botvinick, 2012; Pezzulo et al., 2013) and *active inference* (Friston, 2010; Clark, 2013), where they have a key role in guiding action selection and control (Lepora and Pezzulo, 2015, Pezzulo and Castelfranchi, 2009, Pezzulo et al., 2014a, Pezzulo et al., 2014b, Verschure et al., 2014). Our model complements these proposals by offering a mechanistic explanation of how their required goal representations might be learned in the first place. Furthermore, our results suggest that encoding the prospective goal might simplify cognitive control tasks by permitting splitting them into two distinct phases, goal identification and target selection, and to carry on only limited information (the target identity) from the former to the latter. In this perspective, an advance representation of the identity of the to-be-selected target stimulus is an efficient way to encode context information in cognitive control tasks, and couples *accuracy* (it permits the model, or the monkey, to respond adequately when the target appears) and *parsimony* (only the identity of the target stimulus need to be remembered, not its other features such as its magnitude).

The results of our study parallel a body of evidence in human neuroscience that shows the relevance of nonparametric methods to understand human learning and cognitive control (Collins & Frank 2013; Collins & Koechlin 2012; Donoso et al., 2014). Reassuringly, all these complementary research streams show that the same set of computational methods can apply to a variety of data obtained using different techniques, single cell neurophysiological responses in monkeys, and human fMRI or EEG data. The convergence of results in these computational studies suggest that some of the benefits of the nonparametric model, such as its usefulness for transfer learning, might have general application, as they have been reported in previous simulations (Collins & Frank 2013) and confirmed here in a very different set-up.

Our study also points to the importance of using appropriate state or task representations for solving cognitive control tasks. The Wilson et al. (2014) study established a role for the orbitofrontal cortex in state representations, but did not address the problem of how to learn them. Here, instead, we discuss the (nonparametric) computational mechanisms that permit learning state representations that encode prospective goals and mapping these mechanisms to single cell properties of the monkey prefrontal cortex.

We verified the robustness, generality, and scalability of the obtained results using various control learning simulations. First, we showed that the result is robust with respect to the level of perceptual noise. To this aim we applied the same behavioral protocol of the main simulation but we introduced either half or double perceptual noise. We found that accuracy correspondingly increased

or decreased, as expected, but the same goal-coding principle emerged (Figure 6a-f). We then assessed whether the goal-coding we found was specific to the block-design task presentation or if it also emerged using a more ecologically valid design in which multiple tasks are interleaved. To this aim we applied a behavioral paradigm in which the tasks were presented in entirely interleaved design, i.e., they were pseudo-randomly selected across all learning trials. The results demonstrated that the same goal-coding strategy emerged and guided the behavior demonstrating the generality of the approach (Figure 6g-i). Finally, we verified whether the goal-coding principle would scale beyond simple dichotomous choices. We thus designed a more challenging task with multiple possible goals in which the identity (i.e., color) of context stimuli S1 and S2 was pseudo-randomly selected among k colors, and all the k colors were presented in random order as target stimuli. Simulations with three and four target colors resulted in successful learning and the analyses revealed that also in these more challenging situations the behavior was guided by emergent goals (Figure 8).

Overall, the control simulations indicate that the goal-coding principle extends beyond the specific conditions of our reference monkey neurophysiological study (Genovesio et al., 2012) and applies to various more challenging conditions, demonstrating the scalability of the non-parametric value learning approach to situations that include stimulus-response-value contingencies that are very noisy and presented in variable order. For example, the large-noise control simulations explain how infants with not fully developed perceptual system could nevertheless robustly extract implicit goals despite very noisy internal stimuli representations (e.g., Feigenson, 2011). More generally, the control simulations correspond to novel empirical predictions that remain to be tested by future research.

The current model has also some limitations, and in particular it eschews the full complexity of PFC responses in cognitive control tasks. For example, Genovesio et al, 2014 report that, in one of the tasks studied here, PFC neurons carry information that is not related to the current trial (e.g., information about past goal and outcomes). This information was irrelevant - in fact, the monkeys were not required to maintain that information in memory to correctly perform the task - and future studies are needed to assess whether this information is used for action selection or other functions such as monitoring. Despite so, this evidence raises the intriguing issue that the carry-on of information from one trial to another might be used to learn the long-term statistics of the task or its try-by-trial structure, which would require an extension of the current model.

The novel analytical-computational approach adopted in this study is of general interest since it applies to a variety of neurophysiological studies. The non-parametric Bayesian approach we used (Gershman and Blei, 2012) affords efficient approximate learning of complex nonlinear latent features within a probabilistic generative framework (Sanborn et al., 2010), providing an excellent vehicle for neural-level analysis alike connectionist neurocomputational modeling (Stoianov and Zorzi, 2012). We framed the proposed learning procedure at a high, so-called "computational" of analysis. However, plausible biological implementations have been proposed for the belief propagation methods that we used for the inference (Doya et al., 2007), along with approximate inference methods that permit addressing larger state spaces (Friston, 2010). The overall nonparametric approach has a viable biological implementation, too, and points to hierarchical statistical learning in prefrontal hierarchies (Friston, 2008; Frank and Badre, 2012; Koechlin and Summerfield, 2007) shaped by reinforcement-related signals through prefrontal – (ventral) basal ganglia loops (O'Reilly and Frank, 2006). Collins and Frank (2013) showed that the non-parametric Dirichlet process used here can be neurally implemented and has good quantitative fits to the behaviour produced by a neural network (in which the sparseness of the connectivity matrix from

contexts to PFC was linked to the alpha clustering parameter), even though the mapping is not exact. Exploring the detailed biological mechanisms underlying the proposed nonparametric model is an open objective for future research.

SUMMARY AND CONCLUSION

We report a computational study suggesting that goal-coding at the single cell level represents an efficient solution to cognitive control problems: it permits selecting among the available actions based on the current task and goal contingencies (Miller and Cohen, 2001; Passingham and Wise, 2012), has low memory requirements, and permits to learn faster novel tasks (transfer learning) by aggregating novel unseen contexts to context-categories learned in previous tasks and thus reusing existing sensory-motor strategies (Figure 2a, 6). Goal-coding might be a fundamental organizing principle of prefrontal cortex (Koechlin et al., 2003; Passingham and Wise, 2012), analogous to efficient coding principles in other (e.g. visual) brain areas, and one of its neural signatures might be the modulation of PFC tuning profiles depending on task-relevant rules (Stokes et al., 2013).

REFERENCES

Anderson JR (1991) The adaptive nature of human categorization. *Psychol Rev 98*:409–429.

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition 113*:262–280.

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci 36*:181–204.

Collins AGE, Frank MJ (2013) Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol Rev 120*:190–229.

Collins AGE, Koechlin E (2012) Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol 10*:e1001293.

Dayan P (2009) Goal-directed control and its antipodes. *Neural networks 22*:213–219.

Dolan RJ, Dayan P (2013) Goals and Habits in the Brain. *Neuron 80*:312–325.

Donoso M, Collins a. GE, Koechlin E (2014) Foundations of human reasoning in the prefrontal cortex. *Science (80-)* 1481.

Doya K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw 12*:961–974.

Doya K, Ishii S, Pouget A, Rao RPN (2007) Bayesian Brain: Probabilistic Approaches to Neural Coding (Doya K, Ishii S, Pouget A, Rao RPN, eds). The MIT Press.

Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci 2*:820–829.

Feigenson, L. (2011). Predicting sights from sounds: 6-month-olds' intermodal numerical abilities. *Journal of Experimental Child Psychology, 110(3)*, 347–61.

Frank MJ, Badre D (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb cortex 22*:509–526.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science 291*:312–316.

Friston K (2008) Hierarchical models in the brain. PLoS Comput Biol 4:e1000211.

Friston K (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci 11*:127–138.

Friston K, Schwartenbeck P, Fitzgerald T, Moutoussis M, Behrens T, Dolan RJ (2013) The anatomy of choice: active inference and agency. *Front Hum Neurosci 7*:598.

Fuster JM (1997) The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe. Lippincott–Raven.

Genovesio A, Brasted PJ, Wise SP (2006) Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. *J Neurosci 26*:7305–7316.

Genovesio, A., Tsujimoto, S., Navarra, G., Falcone, R., Wise, S.P. (2014) Autonomous encoding of irrelevant goals and outcomes by prefrontal cortex neurons. Journal of Neuroscience 34, 1970–1978.

Genovesio A, Tsujimoto S, Wise SP (2012) Encoding goals but not abstract magnitude in the primate prefrontal cortex. *Neuron 74*:656–662.

Gershman SJ, Blei DM (2012) A tutorial on Bayesian nonparametric models. *J Math Psychol 56*:1–12.

Ghahramani Z (2013) Bayesian non-parametrics and the probabilistic approach to modelling. *Philos Trans A Math Phys Eng Sci*.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psycho-logical Review, 84*, 279–325.

Ince R A A, Mazzoni A, Bartels A, Logothetis NK, Panzeri S (2012) A novel test to determine the significance of neural selectivity to single and multiple potentially correlated stimulus features. *J Neurosci Methods 210*:49–65.

Koechlin E, Hyafil A (2007) Anterior prefrontal function and the limits of human decision-making. *Science 318*:594–598.

Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. *Science 302*:1181–1185.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in cognitive sciences*, 11(6), 229-235.

Legenstein, R., & Maass, W. (2014). Ensembles of Spiking Neurons with Noise Support Optimal Probabilistic Inference in a Dynamically Changing Environment. *PLoS computational biology, 10(10)*, e1003859. Chicago

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences, 7(4)*, 145-147.

Lepora, N, Pezzulo, G (2015) Embodied Choice: How action influences perceptual decision making. *PLOS Computational Biology* 11(4): e1004110

Miller EK, Cohen J (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci 24*:167–202.

Monsell S (2003) Task switching. *Trends Cogn Sci 7*:134–140.

O'Reilly R, Herd S, Pauli W (2010) Computational models of cognitive control. *Curr Opin Neurobiol 20*:257–261.

O'Reilly RC, Frank MJ (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput 18*:283–328.

Panzeri S, Senatore R, Montemurro M a, Petersen RS (2007) Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol 98*:1064–1072.

Passingham R, Wise S (2012) The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight. OUP Oxford.

Pezzulo G (2012) An Active Inference view of cognitive control. *Front Psychol* 3:478.

Pezzulo, G, Castelfranchi, C (2009). Thinking as the control of imagination: a conceptual framework for goal-directed systems. *Psychological Research*, 73(4):559–577.

Pezzulo G, Rigoli F, Chersi F (2013) The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front Psychol 4*:92.

Pezzulo G, van der Meer MA, Lansink CS, Pennartz CM (2014a) Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn Sci 18*:647–657.

Pezzulo G, Verschure PFMJ, Balkenius C, Pennartz CM (2014b) The principles of goal-directed decision-making: from neural mechanisms to computation and robotics. *Philos Trans R Soc Lond B Biol Sci 369*.

Pica P, Lemer C, Izard V, Dehaene S (2004) Exact and approximate arithmetic in an Amazonian indigene group. *Science 306*:499–503.

Pouget A, Beck JM, Ma WJ, Latham PE (2013) Probabilistic brains: knowns and unknowns. *Nat Neurosci 16*:1170–1178.

Reverberi C, Görgen K, Haynes J-D (2012) Compositionality of rule representations in human prefrontal cortex. *Cereb cortex 22*:1237–1246.

Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature 497*:585–590.

Sanborn AN, Griffiths TL, Navarro D (2010) Rational approximations to rational models: alternative algorithms for category learning. *Psychol Rev 117*:1144–1167.

Seger C, Miller EK (2010) Category learning in the brain. *Annu Rev Neurosci 33*:203–219.

Shima K, Isoda M, Mushiake H, Tanji J (2007) Categorization of behavioural sequences in the prefrontal cortex. *Nature 445*:315–318.

Solway A, Botvinick MM (2012) Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological Review 119*:120–154.

Stoianov I, Zorzi M (2012) Emergence of a "visual number sense" in hierarchical generative models. *Nat Neurosci* 15:194–196.

Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron 78*:364–375.

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.

Tudusciuc O, Nieder A (2007) Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex. *Proc Natl Acad Sci USA*, *104*:14513–14518.

Verschure, P., Pennartz, C. M. A., & Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos Trans R Soc Lond B Biol Sci*, 369: 20130483.

Whalen J, Gallistel CR, Gelman R (1999) Nonverbal Counting in Humans: The Psychophysics of Number Representation. *Psychological Science* 10:130–137.

Yamagata T, Nakayama Y, Tanji J, Hoshi E (2012) Distinct information representation and processing for goal-directed behavior in the dorsolateral and ventrolateral prefrontal cortex and the dorsal premotor cortex. *J Neurosci* 32:12934–12949.