PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Research

Cite this article: Koechlin E. 2014 An evolutionary computational theory of prefrontal executive function in decision-making. *Phil. Trans. R. Soc. B* **369**: 20130474.
http://dx.doi.org/10.1098/rstb.2013.0474

One contribution of 18 to a Theme Issue 'The principles of goal-directed decision-making: from neural mechanisms to computation and robotics'.

Subject Areas:
cognition, computational biology, neuroscience, behaviour

Keywords:
prefrontal cortex, executive control, decision-making, reasoning, Bayesian inference, reinforcement learning

Author for correspondence:
Etienne Koechlin
e-mail: etienne.koechlin@upmc.fr

An evolutionary computational theory of prefrontal executive function in decision-making

Etienne Koechlin

Institut National de la Santé et de la Recherche Médicale, Université Pierre et Marie Curie,
Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

The prefrontal cortex subserves executive control and decision-making, that is, the coordination and selection of thoughts and actions in the service of adaptive behaviour. We present here a computational theory describing the evolution of the prefrontal cortex from rodents to humans as gradually adding new inferential Bayesian capabilities for dealing with a computationally intractable decision problem: exploring and learning new behavioural strategies versus exploiting and adjusting previously learned ones through reinforcement learning (RL). We provide a principled account identifying three inferential steps optimizing this arbitration through the emergence of (i) factual reactive inferences in paralimbic prefrontal regions in rodents; (ii) factual proactive inferences in lateral prefrontal regions in primates and (iii) counterfactual reactive and proactive inferences in human frontopolar regions. The theory clarifies the integration of model-free and model-based RL through the notion of strategy creation. The theory also shows that counterfactual inferences in humans yield to the notion of hypothesis testing, a critical reasoning ability for approximating optimal adaptive processes and presumably endowing humans with a qualitative evolutionary advantage in adaptive behaviour.

1. Introduction

The prefrontal cortex subserves executive control and decision-making for coordinating and selecting thoughts and actions in the service of adaptive behaviour. Present in all mammals [1], the prefrontal cortex in rodents mainly reduces to paralimbic brain regions including the orbitofrontal cortex (OFC) and anterior-cingulate cortex (ACC) [1]. In primates, the prefrontal cortex has evolved with the development of lateral prefrontal regions (LPC) [2]. In humans, the LPC has further evolved with the emergence of the left–right asymmetry yielding to the notion of Broca's area [3,4] subserving human language [5] and bilaterally, in its most anterior portion, a polar region [6,7] (lateral frontopolar cortex, lFPC) which apparently has no homologues in monkeys [8,9] and subserves human reasoning [10].

The prefrontal cortex forms loop circuits with basal ganglia. These subcortical brain nuclei are common to vertebrates and include especially the striatum, which subserves reinforcement learning (RL) [11–14]. RL and, more specifically, temporal-difference RL algorithms are basic online adaptive processes that adjust a behavioural strategy mapping stimuli onto actions according to the discrepancy between actual and expected rewards. Importantly, RL is both a very simple and robust adaptive process that can learn a variety of complex tasks even in uncertain environments. In particular, when rewards only depend upon current states and actions and each state is encountered sufficiently often, RL converges towards the behavioural strategy maximizing rewards [15]. Evidence in rodents, primates and humans indicates that the ventral striatum processes reinforcing signals such as reward prediction errors that serve to adjust stimulus–response associations, whereas the dorsal striatum in

Royal Society Publishing

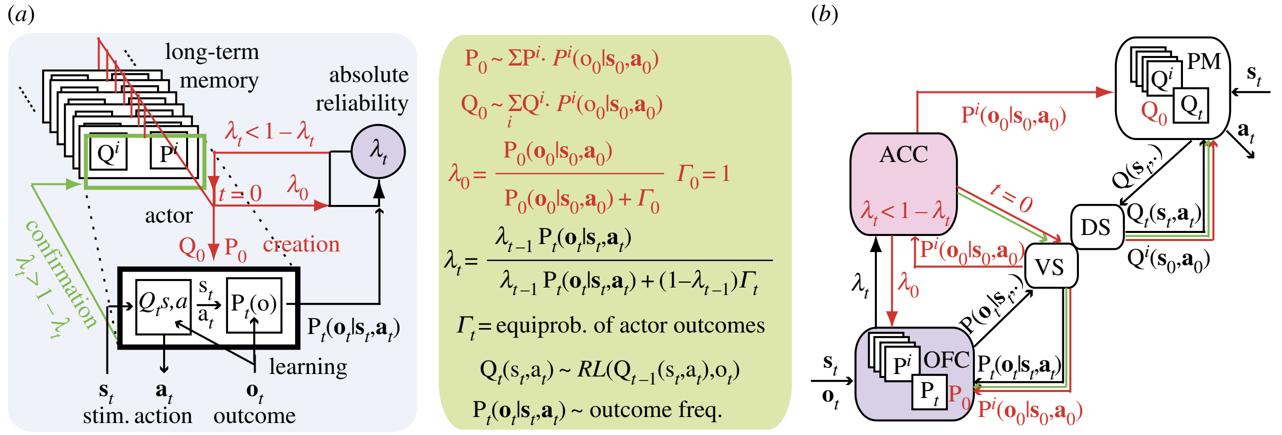
(a)
long-term memory
absolute reliability
Q^i
P^i
$\lambda_t < 1-\lambda_t$
λ_t
$t = 0$
λ_0
actor
Q_0 P_0
creation
confirmation $\lambda_t > 1-\lambda_t$
$Q_t s,a$
s_t a_t
$P_t(o)$
$P_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t)$
learning
$\mathbf{s}_t$ $\mathbf{a}_t$ $\mathbf{o}_t$
stim. action outcome
$P_0 \sim \Sigma P^i \cdot P^i(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)$
$Q_0 \sim \sum_i Q^i \cdot P^i(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)$
$\lambda_0 = \frac{P_0(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)}{P_0(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)+\Gamma_0}$ $\Gamma_0 = 1$
$\lambda_t = \frac{\lambda_{t-1} P_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t)}{\lambda_{t-1} P_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t)+(1-\lambda_{t-1})\Gamma_t}$
Γ_t = equiprob. of actor outcomes
$Q_t(s_t,a_t) \sim RL(Q_{t-1}(s_t,a_t),o_t)$
$P_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t) \sim$ outcome freq.
(b)
PM
Q^i
Q_0
Q_t
$\mathbf{s}_t$
$\mathbf{a}_t$
$P^i(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)$
ACC
$\lambda_t < 1-\lambda_t$
$t = 0$
$Q(\mathbf{s}_t,.)$
$Q_t(\mathbf{s}_t,\mathbf{a}_t)$
DS
$Q^i(\mathbf{s}_0,\mathbf{a}_0)$
$P^i(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)$
VS
λ_t
λ_0
$P(\mathbf{o}|\mathbf{s}_t,.)$
P^i
OFC
P_t
P_0
$P_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t)$
$P^i(\mathbf{o}_0|\mathbf{s}_0,\mathbf{a}_0)$
$\mathbf{s}_t$
$\mathbf{o}_t$

Figure 1. Factual reactive inferences in the rodent prefrontal cortex. (*a*) Inferential system arbitrating between actor learning and creation from long-term memory. Q, selective and P, predictive models forming behavioural strategies stored in the long-term memory repertoire (superscript ***i***). Subscript ***t*** (trial number) indicates actor strategy driving ongoing behaviour and learning external contingencies through reinforcement learning (RL) and action outcome frequencies. λ_t, actor absolute reliability inferred online (right inset: $\boldsymbol{\Gamma}_t$, default-likelihood). $t = 0$, time when the actor becomes unreliable ($\lambda_t < 1 - \lambda_t$) and a new actor is created by mixing stored strategies weighted according to predictive models. Green, confirmation events when the actor becomes reliable ($\lambda_t > 1 - \lambda_t$). (*b*) Presumed system implementation in paralimbic prefrontal regions (rodents). DS, dorsal and VS, ventral striatum filtering out non-actor strategies and learning external contingencies. PM, ACC and OFC, premotor, anterior-cingulate cortex and orbitofrontal cortex, respectively. Red, actor creation triggered in ACC; actor filtering in the striatum is off and allows mixing strategies stored in OFC and PM. See text for explanation.

relation to the premotor cortex processes stimulus–response associations guiding action selection [13,16–18].

However, RL has severe adaptive limitations. The most evident and crucial limitation is that learning new behavioural strategies erases previously learned ones. Indeed, the ability to store and re-use previously learned strategies confers an evolutionary advantage in environments exhibiting external contingencies that change and reoccur periodically (i.e. recurrent situations). In open-ended environments, however, where additionally new external contingencies may always appear, arbitrating between exploring/learning new behavioural strategies versus exploiting/adjusting previously learned ones raises an intractable computational problem. Here, we propose a computational theory postulating that the prefrontal cortex has evolved as primary solving this arbitration problem.

The statistically optimal solution involving Dirichlet processes mixtures [19] is computationally intractable for the two following reasons. First, arbitrating between creating new strategies versus adjusting previously learned ones is in essence non-parametric. This requires optimal adaptive processes to systematically re-evaluate offline past arbitrations whenever new information is acquired and consequently, to revise the repertoire of previously learned strategies in a backward fashion. Second, optimal adaptive processes require monitoring online the whole repertoire of learned strategies that continuously increase when new strategies are created. These computational requirements for optimal arbitrations rapidly yield to intractable computations, suggesting that the prefrontal cortex has evolved as implementing online only forward inferences over a small portion of the repertoire of learned strategies. Accordingly, our theory postulates that under these constraints, the development of prefrontal regions from lower mammalians to humans gradually adds new inferential/computational capabilities beyond RL, which increasingly optimize the arbitration between exploring/learning new behavioural strategies versus exploiting/adjusting previously learned ones. Assuming that basal ganglia implement the ongoing strategy that adjusts to external contingencies through RL and guides behaviour (referred to as *the actor* strategy or simply *the actor*), the prefrontal cortex may have evolved as arbitrating online between these two options: (i) staying with the current actor strategy, which adjusts through RL to external contingencies and (ii) switching away from the current actor strategy and creating a new one from the previously learned strategies stored in long-term memory for driving subsequent behaviour.

2. Reactive inferences, memory recollection and the paralimbic prefrontal cortex

Arbitrating between these two options first requires inferring when external contingencies change and require switching away from the current actor strategy. Our theory assumes that the development of paralimbic prefrontal regions in lower mammals implements a first inferential step: namely, inferring such changes based on the inconsistency between actual action outcomes and the outcome contingencies the actor has learned so far. Such inferences are *factual* as they only bear upon the outcome predictive model the actor strategy has learned, and *reactive* as they operate only after observing action outcomes.

To make such inferences, the actor thus learns both a *selective* and *predictive* model: the former maps stimuli onto actions, adjusts through RL and enables selection of the most rewarding actions in response to stimuli; the latter maps stimulus–action associations onto expected outcomes and learns by simply registering outcome frequencies given responses to stimuli. Critically, the predictive model enables inference online of the actor *absolute reliability*, i.e. the posterior probability the current external contingencies match those the actor has learned (figure 1*a*). Updating online actor absolute reliability according to actual action outcomes involves forward Bayesian inferences and requires comparison of the likelihood of actual action outcomes derived from the actor predictive model with their likelihood according to *any* potential alternative models. The latter cannot be exactly computed, because the range of possible external contingencies is presumably infinite and unknown. However, that likelihood can be estimated as reflecting the maximal predictive entropy, namely as the equiprobability of action outcomes *produced by* the actor. We referred to this estimate as the *default outcome-likelihood*.

Actor absolute reliability λ_t in every trial t serves to arbitrate between staying versus switching from the current actor strategy. When the actor remains more likely reliable than unreliable ($\lambda_t > 1 - \lambda_t$), no changes in external contingencies are likely to have occurred. The same actor strategy is then kept and continues to adjust through RL. The system thus operates in an *exploitation* mode. When conversely the actor becomes unreliable ($\lambda_t < 1 - \lambda_t$), external contingencies have likely changed. A new actor is then built from optimally using the whole repertoire of previously learned strategies stored in long-term memory, which correspond to the former actor strategies (including the one that has just become unreliable). The new actor selective (predictive, resp.) model is thus formed as the mixture of selective (predictive, resp.) models stored in the repertoire possibly weighted according to current action outcome given strategies' predictive models (figure 1*a*).

In this mixture process, importantly, selective models may be recalibrated according to current rewarding values of action outcomes through a *model-based* RL process [20]: selective and predictive models serve as action and outcome predictors, respectively, for implementing covert RL, whereby outcome rewarding values are possibly altered with respect to current animal needs (e.g. satiety effects). This model-based RL process calibrates the new actor selective model according to current animal needs.

As the new actor is created from the strategies repertoire, its initial absolute reliability corresponds to the repertoire absolute reliability, i.e. the probability the current external contingencies match those associated with one stored strategy, or equivalently, the probability that the animal faces a previously encountered situation. This absolute reliability is inferred using Bayes' law, which requires evaluating the likelihood of the current action outcome in every previously encountered situation and separately, in any possible new situations (figure 1, inset). In the former case, this likelihood is simply derived from the mixture of stored predictive models (corresponding to the new actor predictive model). In the latter case, the likelihood is estimated as the default outcome-likelihood associated with the presumably new situation. It is actually equal to 1, because in this situation, only one outcome has been observed. As a result, the new actor guiding behaviour is initially inferred as being unreliable (figure 1, inset). The system thus operates in an *exploration* mode that promotes actor learning by preventing switching again, *while* the actor remains unreliable. When the actor becomes reliable, the system returns to the exploitation mode, whereby switching away from the actor strategy (when it again becomes unreliable) and creating a new one may occur again. Note that the lower the initial actor reliability is, the longer exploration will last: initially, the new actor less likely matches the new situation and, consistently, more trials are required for the new actor to learn the new external contingencies.

Under its intrinsic computational constraints (forward, factual and reactive inferences only), this model is an optimal adaptive system in environments featuring both new and recurrent situations. The model especially exhibits three key functional properties, which are consistent with empirical data. First, the model shows abrupt rather than gradual behavioural changes, when following variations in external contingencies, the actor strategy becomes unreliable, and a new one is created. Such abrupt behavioural changes are routinely observed in rodents, primates and human experiments [21–24]. Second, the long-term repertoire of behavioural strategies expands whenever new actors are created, so that the *reoccurrence* frequencies of external situations have a major influence on shaping new actors. Whenever external situations reoccur, new actors are created with selective and predictive models learning again the associated external contingencies, thereby replicating in the repertoire the selective and predictive models previously learned from previous occurrences: the more external situations reoccur, the more these models are then replicated in long-term memory. Consequently, external contingencies learned from situations that more frequently reoccur contribute more to the formation of new actor strategies. This computational model thus exhibits a basic feature of Dirichlet processes [19]. Collins & Koechlin [25] showed that the model accounts for increasing human performances associated with recurrent situations. Third, actor creation involves model-based RL, whenever ongoing actors driven by standard (model-free) RL become unreliable. Arbitrating between adjusting the current actor versus creating a new one thus yields to a decision between model-based and model-free RL, which accounts for behavioural changes observed in rodents (e.g. extinction effects) following outcome devaluation manipulations [20].

As generally agreed, the premotor cortex along with the dorsal striatum encodes and stores selective models of behavioural strategies [26–28], whereas the dorsal and ventral striatum implement RL adjusting the actor selective model (see above). Our hypothesis is that operating beyond RL, the factual and reactive inferential system is implemented in paralimbic prefrontal regions (figure 1, right). The OFC (especially the medial OFC in humans) encodes strategies' predictive models and updates actor absolute reliability according to action outcomes. Empirical data in rodents, monkeys and humans show that, consistently, this region responds to action outcomes in relation with outcome predictions [29–33]. Moreover, human neuroimaging studies showed that the medial OFC is involved in monitoring the ongoing course of action [34] and inferring changes in external contingencies [33].

By contrast, the ACC is assumed to detect when the actor becomes unreliable for triggering the creation of new actors. Rodent studies show that in the ACC, neuronal activity consistently exhibits abrupt phase transitions in relation to behavioural switches [24]. Moreover, monkey electrophysiological and human neuroimaging studies indicate that the ACC is involved in monitoring when to switch from exploitation to exploration behaviours [35,36], whereas adjacent medial (pre)supplementary motor regions are involved in inhibiting established behavioural responses [37] and promoting exploratory responses [38]. More generally, the ACC is involved in responding to surprising outcomes triggering behavioural switches [39] and in starting the execution of new tasks [40].

Empirical data (review in [41]) further suggest that prefrontal–striatal loop circuits involving the ACC and OFC [42] may subserve actor creation. The ACC may prevent the striatum from filtering out non-actor strategies and allow stored strategies to mix for forming new actors: the ventral striatum, which for every stored strategy then receives the outcome-likelihood from the OFC, may return these likelihoods to both the ACC and OFC; assuming that the ACC further conveys these pieces of information to the premotor

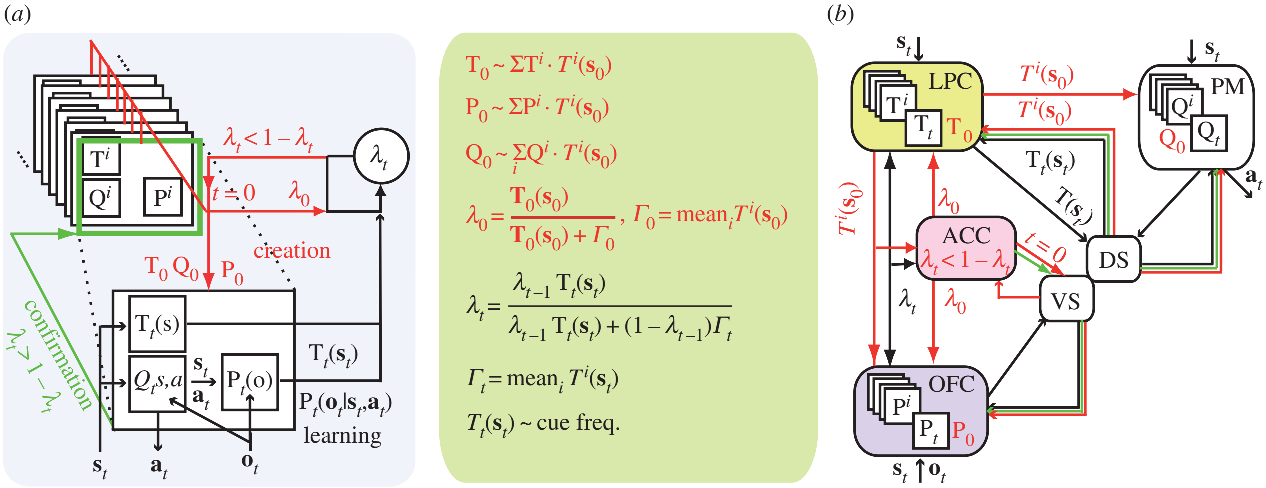
(a)
T^i
Q^i
P^i
$\lambda_t < 1 - \lambda_t$
$t = 0$
λ_0
λ_t
creation
T_0 Q_0 P_0
confirmation
$\lambda_t > 1 - \lambda_t$
$\mathrm{T}_t(\mathrm{s})$
$Q_t s,a$
$\mathbf{s}_t$
$\mathbf{a}_t$
$\mathrm{P}_t(\mathrm{o})$
$\mathrm{T}_t(\mathbf{s}_t)$
$\mathrm{P}_t(\mathbf{o}_t|\mathbf{s}_t,\mathbf{a}_t)$
learning
$\mathbf{s}_t$
$\mathbf{a}_t$
$\mathbf{o}_t$
$\mathrm{T}_0 \sim \Sigma \mathrm{T}^i \cdot T^i(\mathbf{s}_0)$
$\mathrm{P}_0 \sim \Sigma \mathrm{P}^i \cdot T^i(\mathbf{s}_0)$
$\mathrm{Q}_0 \sim \sum_i \mathrm{Q}^i \cdot T^i(\mathbf{s}_0)$
$\lambda_0 = \frac{\mathbf{T}_0(\mathbf{s}_0)}{\mathbf{T}_0(\mathbf{s}_0) + \Gamma_0}, \Gamma_0 = \mathrm{mean}_i T^i(\mathbf{s}_0)$
$\lambda_t = \frac{\lambda_{t-1}\,\mathrm{T}_t(\mathbf{s}_t)}{\lambda_{t-1}\,\mathrm{T}_t(\mathbf{s}_t) + (1 - \lambda_{t-1})\Gamma_t}$
$\Gamma_t = \mathrm{mean}_i\, T^i(\mathbf{s}_t)$
$T_t(\mathbf{s}_t) \sim$ cue freq.
(b)
$\mathbf{s}_t$
LPC
T^i
T_t
T_0
$T^i(\mathbf{s}_0)$
$T^i(\mathbf{s}_0)$
$\mathbf{s}_t$
PM
Q^i
Q_0
Q_t
$\mathbf{a}_t$
$\mathrm{T}_t(\mathbf{s}_t)$
$T(\mathbf{s}_t)$
$T^i(\mathbf{s}_0)$
λ_0
ACC
$\lambda_t < 1 - \lambda_t$
$t = 0$
DS
λ_t
λ_0
VS
OFC
P^i
P_t
P_0
$\mathbf{s}_t$ $\mathbf{o}_t$

Figure 2. Factual reactive and proactive inferences in the primate prefrontal cortex. (*a*) Same inferential system shown in figure 1 but adding proactive inferences arbitrating between actor learning and creation according to current external cues (red, actor creation following proactive inferences). T, contextual models learning likelihoods of external cues associated with strategies and allowing revising actor absolute reliability before acting (inset). (*b*) Presumed implementation of proactive inferences in lateral prefrontal regions (LPC, primates) in addition to reactive inferences in OFC shown in figure 1. Detailed legend in figure 1.

cortex, the mixture of selective and predictive models given current action outcomes may then occur in the premotor cortex and OFC, respectively. The ACC may concomitantly initialize actor reliability.

The theory indicates that new actors are initially unreliable (exploration mode) and when they become reliable, the system returns to the exploitation mode. This event may correspond to an internal reinforcing signal consolidating new actor selective and predictive models in long-term memory. This event may thus be detected in the ventral striatum, which processes behavioural reinforcers and receives projections from the OFC presumably inferring actor reliability. The theory thus predicts that the ACC triggers exploration, whereas the ventral striatum signals when to return to exploitation.

3. Proactive inferences and contextual control in the lateral prefrontal cortex

The above-described inferential system has two critical limitations. First, its adaptive capability is only reactive: new actor strategies are created only after experiencing action outcomes, which may be detrimental with adverse outcomes. Second, actor creation ignores the context in which stored strategies were learned. Accordingly, our theory assumes that the development of LPCs in primates implements a second inferential step overcoming these limitations: namely, further inferring from external cues when to switch away from the actor strategy. In contrast to action outcomes, external cues occur independently of subjects' behaviour but their occurrences may also inform about changes in external contingencies. The resulting inferential system thus exhibits *proactive* behaviours, because external cues typically alter the arbitration between adjusting versus creating actor strategies before acting.

For making such proactive inferences, the actor strategy learns an additional internal model, which we refer to as the *contextual model*. The actor contextual model simply registers the frequencies of external cues and is stored in long-term memory along with selective and predictive models (figure 2). When external cues occur, the actor absolute reliability is then updated through forward Bayesian inference, which requires comparison of the likelihood of these external cues derived from the actor contextual model with their likelihood according to *any* potential alternative models. The latter is again not exactly computable. External cues are however independent of subjects' behaviour. This likelihood can therefore be estimated as the frequency of current external cues observed in the past, which simply derives from the mean of contextual models stored in long-term memory. This estimate is referred to as the *default context-likelihood*. Consequently, external cues that are less likely to occur in the current than past situations, degrade the actor absolute reliability and may yield to proactively switch away from the current actor strategy for creating a new one.

Whenever new actors are created following reactive and/or proactive inferences, their selective (predictive and contextual, respectively) model is again computed as the mixture of selective (predictive and contextual, resp.) models stored in long-term memory. The mixture however is now weighted according to current action outcomes and/or external cues given predictive and/or contextual models (figure 2). Accordingly, actor creation may now depend upon current external cues along with action outcomes and the frequency of recurrent situations. In particular, strategies learned within more distinct contexts than the current one contribute less to actor creation.

New actor initial reliability again corresponds to the repertoire absolute reliability, but the latter is now evaluated according to current action outcomes and external cues (figure 2, inset). This absolute reliability is computed as above based on new actor predictive and contextual models in relation to the default outcome- and context-likelihood. Importantly, new actor strategies may now be formed as being immediately reliable; this happens when current external cues match those under which strategies already in the repertoire were learned. In that event, new actor strategies may then be rejected as soon as they serve as actor according to subsequent cue- and outcome-based counterevidence. Proactive inferences thus provide the ability to control behaviour according to the context, i.e. to rapidly recreate and switch across behavioural strategies according to external cues.

Under its intrinsic computational constraints (forward and factual inferences only), this computational model optimally uses external cues and action outcomes for adapting to environments featuring both new and recurrent situations. Our hypothesis is that the lateral prefrontal cortex (LPC) learns and encodes contextual models and updates the actor absolute reliability according to external cues (figure 2). Monkey and human studies show that, consistently, the LPC subserves the formation and selection of behavioural strategies according to contextual cues [43–45]. The neuronal connections between the LPC, OFC and premotor cortex [46] are further assumed to link the contextual, predictive and selective models associated with the same strategy. More specifically, the LPC revises the actor absolute reliability conveyed from the OFC according to external cues and returns the updated reliability to the OFC (and vice versa). The ACC again detects when the actor becomes unreliable for triggering actor creation.

Actor creation requires mixing of selective (predictive and contextual, resp.) models over stored strategies according to outcome and cue likelihoods. This may be achieved as described above through the architecture of cortical–cortical and striatal–cortical connections within the frontal lobes [42,46] (figure 2). The ACC prevents the striatum from filtering out non-actor strategies, so that for every strategy, the ventral striatum returns the outcome-likelihood to the OFC and ACC, whereas the dorsal striatum returns the cue likelihood to the LPC. The new actor predictive model requires mixing of predictive models, which may occur in the OFC through LPC-to-OFC projections conveying cue likelihoods. The new actor contextual model requires mixing of contextual models, which may occur in the LPC through ACC-to-LPC projections conveying outcome likelihoods. Finally, the new actor selective model requires mixing of selective models, which may occur in the premotor cortex through LPC-to-premotor projections conveying both outcome and cue likelihoods.

4. Counterfactual inferences, hypothesis testing and the frontopolar cortex

In the above-described inferential system, the critical limitation is that inferences remain factual: the decision to adjust versus change the ongoing behavioural strategy, i.e. *the actor*, bears upon the actor reliability only. Accordingly, the theory assumes that the development of the frontopolar cortex (lFPC) in humans implements a third inferential step overcoming this limitation: namely, inferring when to change the current actor strategy from concurrently monitoring the reliability of multiple behavioural strategies. The human executive system thus develops counterfactual inferences bearing upon alternative behavioural strategies that are not guiding ongoing behaviour. These counterfactual inferences enable the inference online not only of *when* to change the actor strategy, but also *which* strategy may replace it. Ideally, counterfactual inferences should bear upon the whole repertoire of stored strategies. This seems however computationally costly and biologically implausible. Our theory therefore assumes that counterfactual inferences develop only over a limited number of stored strategies, referred to as the inferential buffer.

One might consider the inferential buffer as forming a global actor strategy, whereby action selection and strategy learning result from mixing online monitored strategies over the buffer according to their relative reliability [47]. Collins & Koechlin [25] showed that this view is inconsistent with human behavioural performances in sequential decision tasks. This is also theoretically problematic, because the global actor may be inferred as being reliable with only unreliable strategies: the mixture of monitored strategies may therefore be strongly suboptimal when another strategy stored in long-term memory is potentially reliable. More optimally, the executive system may concurrently infer the absolute reliability of every monitored strategy and when none are inferred as being reliable, a new strategy is created from long-term memory (as described above) and added to the inferential buffer. When, conversely, one is inferred as being reliable, the others are necessary unreliable, even when considered collectively (absolute reliabilities sum up to 1 or less). Accordingly, the reliable strategy becomes *the* actor strategy selecting actions and learning external contingencies (i.e. adjusting its selective, predictive and contextual models). The buffer is thus assumed to comprise the actor strategy driving behaviour plus a number of alternative strategies, which for clarity we refer to as *counterfactual* strategies.

The actor strategy may thus be changed rather than adjusted through RL by either *retrieving* and *switching to* a reliable counterfactual strategy or *creating* a new strategy from long-term memory (figure 3*a*). In the former case, the system operates in the *exploitation* mode, because the new actor remains reliable. In the latter case, the new actor may be created as being unreliable; in that case the inferential system switches into the *exploration* mode. The system subsequently returns to the exploitation mode in two ways. Either a counterfactual strategy becomes reliable, while the newly created actor remains unreliable. The former then becomes the actor, and the latter is *rejected* from the buffer and disbanded. Or the newly created actor becomes reliable, whereas the counterfactual strategies remain unreliable. The former is then *confirmed* and stored in long-term memory along with others. Exploration periods thus correspond to *hypothesis testing* on strategy creation. Accordingly, counterfactual strategies are the former actors that have been reliably assigned to an external situation that previously occurred. When newly created actors are confirmed, however, the number of strategies monitored in the buffer increases and possibly reaches its capacity limit. In that event, the theory assumes that the strategy used the least recently as actor is discarded from the buffer. The rationale is that older situations are less frequent and less likely to reoccur in the short-run. The buffer therefore keeps monitoring counterfactual strategies which will more likely match the *next* external situation.

The computations implementing this counterfactual inferential system are essentially the same as those described above. Reactive and proactive inferences are simply extended to counterfactual strategies. The differences are as follows (figure 3, inset): first, absolute reliability is inferred for the actor and counterfactual strategies, so that the *default outcome-likelihood* is now better estimated as the equiprobability of outcomes produced by the actor *and* counterfactual strategies. Second, the absolute reliability of actor and counterfactual strategies directly weights their relative contribution in creating new actors from long-term memory. Collins & Koechlin [25] showed that this computational algorithm predicts human choices in recurrent and new situations featuring uncertain and variable contingencies possibly associated with the occurrences of contextual cues. Moreover, all the model components appeared necessary for accounting for human performances. The best

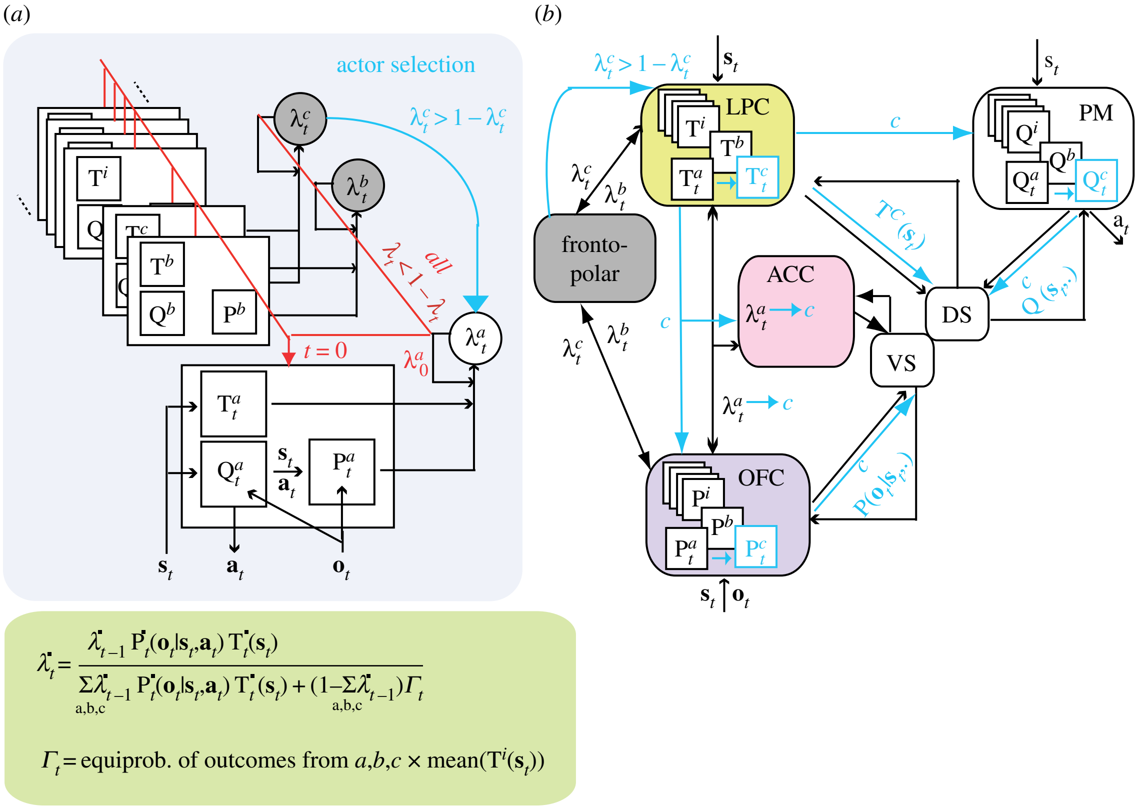
(a)
actor selection
λ^c_t
$\lambda^c_t > 1-\lambda^c_t$
λ^b_t
T^i
T^c
T^b
Q^b
P^b
all $\lambda_t < 1-\lambda_t$
$t = 0$
λ^a_0
λ^a_t
T^a_t
Q^a_t
s_t
a_t
P^a_t
s_t
a_t
o_t
(b)
$\lambda^c_t > 1-\lambda^c_t$
s_t
LPC
T^i
T^b
T^a_t
T^c_t
c
s_t
PM
Q^i
Q^b
Q^a_t
Q^c_t
a_t
λ^c_t
λ^b_t
fronto-polar
$T^C(s_t)$
ACC
$\lambda^a_t \rightarrow c$
c
DS
$Q^c(s_t,.)$
λ^c_t
λ^b_t
VS
$\lambda^a_t \rightarrow c$
OFC
P^i
P^b
P^a_t
P^c_t
c
$P(o_t|s_t,.)$
s_t
o_t
$\lambda^{\bullet}_t = \frac{\lambda^{\bullet}_{t-1} P^{\bullet}_t(o_t|s_t,a_t)\, T^{\bullet}_t(s_t)}{\sum_{a,b,c} \lambda^{\bullet}_{t-1} P^{\bullet}_t(o_t|s_t,a_t)\, T^{\bullet}_t(s_t) + (1-\sum_{a,b,c}\lambda^{\bullet}_{t-1})\Gamma_t}$
Γ_t = equiprob. of outcomes from a,b,c × mean($T^i(s_t)$)

Figure 3. Counterfactual inferences in the human prefrontal cortex. (*a*) Same inferential system shown in figure 2 but adding counterfactual inferences (grey) for arbitrating between actor learning, switching and creation. Superscript a, b, c: actor (a) and two additional counterfactual (b,c) strategies, which absolute reliability is concurrently inferred online (bottom inset, equations show combined proactive and reactive inferences for every monitored strategy). Actor switching (blue, from a to c in the shown example) occurs when one counterfactual strategy becomes reliable. Actor creation (red) occurs when all monitored strategies become unreliable. (*b*) Presumed implementation of counterfactual inferences through the human frontopolar cortex, LPC (proactive component) and OFC (reactive component). The frontopolar cortex encodes counterfactual absolute reliabilities, while the OFC encode actor reliability. Actor switching presumably originates from the LPC and diffuses in the prefrontal network in a top-down fashion (blue arrows). Actor creation is not shown for clarity (identical to figures 1 and 2). Detailed legend in figure 1.

account was found when the buffer capacity corresponds to two/three counterfactual strategies. This size matches the capacity previously proposed for human (declarative) working memory [48].

The hypothesis is that the lFPC encodes the absolute reliability of counterfactual strategies (figure 3*b*). Consistently, neuroimaging studies show that the lFPC is engaged in cognitive branching, that is, holding on the execution of one task during the performance of another task [10,49] and in monitoring the opportunity to switch to alternative courses of action [34]. The lFPC has also major reciprocal connections with the OFC and LPC [9,46]. Accordingly, the (medial) OFC may update counterfactual strategies' reliabilities encoded in the lFPC according to action outcomes (given predictive models presumably stored in the OFC). Similarly, the LPC may update them according to external cues (given contextual models presumably encoded in the LPC).

In contrast to actor creation, retrieving counterfactual strategies as new actors when they become reliable requires *selecting* and reactivating the corresponding selective, predictive and contextual models from long-term memory to drive behaviour. The LPC may be the best candidate for this function (figure 3*b*). Adjacent to the lFPC, the LPC is the only prefrontal region strongly connected to both the premotor cortex and OFC, presumably storing selective and predictive models, respectively [46]. Thus, selecting a reliable contextual model in the LPC may concomitantly induce the selection of associated models in the premotor cortex and OFC. Consistent with the hypothesis, the LPC is involved in retrieving action sets through top-down selection from LPC to premotor regions [44,50].

5. Discussion

We propose here that the prefrontal cortex has primary evolved from lower mammalians to humans by gradually adding new inferential capabilities beyond RL, which progressively optimize the arbitration between exploring/learning new behavioural strategies versus exploiting/adjusting previously learned ones. This arbitration is optimized assuming that the environment varies according to both recurrent and new causes, which are indirectly observable, independent and potentially infinite. Optimization has occurred under the computational constraint that the brain implements only forward, online inferential processes bearing upon a limited portion of behavioural strategies stored in long-term memory.

With these assumptions, optimal arbitration is based on inferring online the *absolute reliability* of every monitored strategy, that is, the posterior probability that given external evidence, the current external contingencies match those the strategy has learned. Monitored strategies may therefore be inferred as being reliable (more likely matching than differing) versus unreliable (the converse). When one is reliable, the others are individually *and* collectively unreliable. We indicate a solution indicating how the brain may compute absolute reliabilities through estimates of *default-likelihoods*, that is, likelihoods of external cues and action outcomes when the current external contingencies presumably match no monitored strategies. The concepts of default-likelihood and absolute reliability generalize the notion of expected/unexpected uncertainty proposed by Yu & Dayan [51] and provide the computational foundations of the present theory. They may be related to the psychological notions of metacognitive processes and confidence judgements [52].

Based on this computational framework, we identify three critical inferential capabilities associated with the development of specific prefrontal regions. The OFC and ACC appearing in rodents provides the ability to make *factual*, *reactive* reliability inferences; reliability inferences are based on action outcomes and only bear upon the actor strategy guiding action and learning external contingencies. The OFC is predicted to encode strategies' internal models predicting action outcomes and to revise actor absolute reliability according to action outcomes. Critically, the actor strategy learns actions through RL so that the more outcomes are rewarding, the more the reliability reflects the predicted occurrence of these outcomes. The ACC, by contrast, detects when the actor strategy becomes unreliable for triggering the creation of new actor strategies from long-term memory. The LPC appearing in primates provides the additional ability to make *factual*, *proactive* inferences. The LPC is predicted to encode strategies' internal models predicting external cues and to revise actor absolute reliability according to external cues typically occurring before action. Finally, the lFPC appearing in humans provides the ability to further make both reactive and proactive *counterfactual* inferences. The lFPC is predicted to encode the absolute reliability of a few counterfactual strategies stored in long-term memory; along with actor reliability, counterfactual reliabilities are revised in OFC and LPC according to action outcomes and external cues, respectively.

Counterfactual strategies along with the actor strategy form an inferential/monitoring buffer essentially equivalent to the psychological notion of procedural working memory [48,53,54]. The notion of actor is consistent with the idea of attentional focus within working memory [53–55]. When one monitored strategy becomes reliable, this strategy becomes the actor. When the actor *becomes* unreliable with no reliable counterfactual alternatives in the buffer, a new actor is created from long-term memory. Actor creation consists of optimally mixing strategies stored in long-term memory according to current outcome-based and/or cue-based evidence. Actor creation is thus a model-based construct of actor strategies involving model-based RL [20]. The theory predicts that this model-based construct of actors occurs through striatal–frontal loop circuits. As newly created actors subsequently adjust through model-free RL, the actor strategy gradually results from the hybridation between model-based and model-free RL, whereby model-free RL progressively dominates with time. This hybridation resulting from abrupt and intermittent model-based constructs of new actors, when model-free RL adjustments of ongoing actors become unreliable, is optimal with forward inferential processes operating online and bearing upon a limited number of counterfactual strategies.

Newly created actors may be inferred as being initially unreliable indicating that the environment is likely in a state that was not previously observed. The prefrontal executive system then switches into an exploration period corresponding to *hypothesis testing*: this unreliable actor guides behaviour and may be subsequently confirmed or rejected. Confirmation occurs when this actor become reliable before any counterfactual strategies. The ventral striatum is predicted to detect this confirmation event yielding to the actor consolidation in long-term memory. Conversely, rejection occurs when a counterfactual strategy become reliable before this actor. The LPC is predicted to retrieve this reliable counterfactual strategy to serve as actor. Critically, hypothesis testing prevents the capacity-limited buffer from monitoring unnecessary strategies emerging from long-term memory. Hypothesis testing, moreover, is a primitive form of backward inferences, because every decision to create new strategies may be subsequently revised on the basis of subsequent information. Backward inferences are actually critical in optimal adaptive systems operating in open-ended environments for dealing with the non-parametric nature of strategy creation [19]. Thus, counterfactual inferences and hypothesis testing associated with the development of the lFPC appear as critical reasoning capabilities endowing humans with a qualitative evolutionary advantage in adaptive behaviour.

The present theory provides a unified, principled model of the overall inferential architecture of the human prefrontal cortex. The model makes testable predictions outlined above [56]. This overall model, however, masks the computational complexity of neuronal processing involved in learning strategies' internal models coding the likelihoods of action outcomes and external cues in the OFC and LPC. Indeed, action outcomes and external cues may vary within high-dimensional, continuous spaces possibly including the time dimension. Accordingly, the OFC and LPC regions are likely to operate in lower-dimensional spaces through categorization and extra/interpolation processes based on relative space metrics [57] for representing action outcomes and external cues and computing their likelihood in relation to behavioural strategies. As recently proposed [57], these coding processes may have concurrently emerged through the development of multiple subregions in OFC and LPC in association with posterior associative cortical regions.

Funding statement. Supported by a European Research Council Grant (ERC-2009-AdG no. 250106).

References

1. Uylings HB, Groenewegen HJ, Kolb B. 2003 Do rats have a prefrontal cortex? *Behav. Brain Res.* **146**, 3–17. (doi:10.1016/j.bbr.2003.09.028)
2. Fuster JM. 1989 *The prefrontal cortex*, 2nd edn. New York, NY: Raven Press.
3. Uylings HB, Jacobsen AM, Zilles K, Amunts K. 2006 Left-right asymmetry in volume and number of neurons in adult Broca's area. *Cortex* **42**, 652–658. (doi:10.1016/S0010-9452(08)70401-5)
4. Schenker NM, Hopkins WD, Spocter MA, Garrison AR, Stimpson CD, Erwin JM, Hof PR, Sherwood CC. 2010 Broca's area homologue in chimpanzees. *Cereb. Cortex* **20**, 730–742. (doi:10.1093/cercor/bhp138)

5. Broca P. 1861 Remarques sur le siège de la faculté du langage articulé suivie d'une observation d'aphémie. *Bull. Soc. Anat.* (*Paris*) **6**, 330.
6. Semendeferi K, Armstrong E, Schleicher A, Zilles K, Van Hoesen GW. 2001 Prefrontal cortex in humans and apes: a comparative study of area 10. *Am. J. Phys. Anthropol.* **114**, 224–241. (doi:10.1002/1096-8644(200103)114:3<224::AID-AJPA1022>3.0.CO;2-I)
7. Teffer K, Semendeferi K. 2012 Human prefrontal cortex: evolution, development, and pathology. *Prog. Brain Res.* **195**, 191–218. (doi:10.1016/B978-0-444-53860-4.00009-X)
8. Koechlin E. 2011 Frontal pole function: what is specifically human? *Trends Cogn. Sci.* **15**, 241. (doi:10.1016/j.tics.2011.04.005)
9. Neubert FX, Mars RB, Thomas AG, Sallet J, Rushworth M. 2014 Comparison of human ventral cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron* **82**, 700–713. (doi:10.1016/j.neuron.2013.11.012)
10. Koechlin E, Hyafil A. 2007 Anterior prefrontal function and the limits of human decision-making. *Science* **318**, 594–598. (doi:10.1126/science.1142995)
11. Stephenson-Jones M, Samuelsson E, Ericsson J, Robertson B, Grillner S. 2011 Evolutionary conservation of the basal ganglia as a common vertebrate mechanism for action selection. *Curr. Biol.* **21**, 1081–1091. (doi:10.1016/j.cub.2011.05.001)
12. Doya K. 2007 Reinforcement learning: computational theory and biological mechanisms. *HFSP J.* **1**, 30–40. (doi:10.2976/1.2732246/10.2976/1)
13. Samejima K, Ueda Y, Doya K, Kimura M. 2005 Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340. (doi:10.1126/science.1115270)
14. Schultz W, Dayan P, Montague PR. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)
15. Sutton RS, Barto AG. 1998 *Reinforcement learning*. Cambridge, MA: MIT Press.
16. Kahnt T, Park SQ, Cohen MX, Beck A, Heinz A, Wrase J. 2009 Dorsal striatal–midbrain connectivity in humans predicts how reinforcements are used to guide decisions. *J. Cogn. Neurosci.* **21**, 1332–1345. (doi:10.1162/jocn.2009.21092)
17. Atallah HE, Lopez-Paniagua D, Rudy JW, O'Reilly RC. 2007 Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat. Neurosci.* **10**, 126–131. (doi:10.1038/nn1817)
18. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. 2004 Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454. (doi:10.1126/science.1094285)
19. Teh YW, Jordan MI, Beal MJ, Blei DM. 2006 Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **101**, 1566–1581. (doi:10.1198/016214506000000302)
20. Daw ND, Niv Y, Dayan P. 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711. (doi:10.1038/nn1560)
21. Asaad WF, Rainer G, Miller EK. 1998 Neural activity in the primate prefrontal cortex during associative learning. *Neuron* **21**, 1399–1407. (doi:10.1016/S0896-6273(00)80658-3)
22. Pasupathy A, Miller EK. 2005 Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* **433**, 873–876. (doi:10.1038/nature03287)
23. Nassar MR, Wilson RC, Heasly B, Gold JI. 2010 An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* **30**, 12 366–12 378. (doi:10.1523/JNEUROSCI.0822-10.2010)
24. Karlsson MP, Tervo DG, Karpova AY. 2012 Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135–139. (doi:10.1126/science.1226518)
25. Collins A, Koechlin E. 2012 Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol.* **10**, e1001293. (doi:10.1371/journal.pbio.1001293).
26. Histed MH, Pasupathy A, Miller EK. 2009 Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. *Neuron* **63**, 244–253. (doi:10.1016/j.neuron.2009.06.019)
27. Hadj-Bouziane F, Meunier M, Boussaoud D. 2003 Conditional visuo-motor learning in primates: a key role for the basal ganglia. *J. Physiol. Paris* **97**, 567–579. (doi:10.1016/j.jphysparis.2004.01.014)
28. Packard MG, Knowlton BJ. 2002 Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* **25**, 563–593. (doi:10.1146/annurev.neuro.25.112701.142937)
29. Jones JL, Esber GR, McDannald MA, Gruber AJ, Hernandez A, Mirenzi A, Schoenbaum G. 2012 Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953–956. (doi:10.1126/science.1227489)
30. McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. 2011 Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705. (doi:10.1523/JNEUROSCI.5499-10.2011)
31. Rudebeck PH, Murray EA. 2011 Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. *J. Neurosci.* **31**, 10 569–10 578. (doi:10.1523/JNEUROSCI.0091-11.2011)
32. Noonan MP, Walton ME, Behrens TEJ, Sallet J, Buckley MJ, Rushworth MFS. 2010 Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl Acad. Sci. USA* **107**, 20 547–20 552. (doi:10.1073/pnas.1012246107)
33. Hampton AN, Bossaerts P, O'Doherty JP. 2006 The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367. (doi:10.1523/JNEUROSCI.1010-06.2006)
34. Boorman ED, Behrens TE, Woolrich MW, Rushworth MF. 2009 How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743. (doi:10.1016/j.neuron.2009.05.014)
35. Quilodran R, Rothe M, Procyk E. 2008 Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* **57**, 314–325. (doi:10.1016/j.neuron.2007.11.031)
36. Kolling N, Behrens TE, Mars RB, Rushworth MF. 2012 Neural mechanisms of foraging. *Science* **336**, 95–98. (doi:10.1126/science.1216930)
37. Isoda M, Hikosaka O. 2007 Switching from automatic to controlled action by monkey medial frontal cortex. *Nat. Neurosci.* **10**, 240–248. (doi:10.1038/nn1830)
38. Donahue CH, Seo H, Lee D. 2013 Cortical signals for rewarded actions and strategic exploration. *Neuron* **80**, 223–234. (doi:10.1016/j.neuron.2013.07.040)
39. Alexander WH, Brown JW. 2011 Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* **14**, 1338–1344. (doi:10.1038/nn.2921)
40. Dosenbach NU *et al*. 2006 A core system for the implementation of task sets. *Neuron* **50**, 799–812. (doi:10.1016/j.neuron.2006.04.031)
41. Liljeholm M, O'Doherty JP. 2012 Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends Cogn. Sci.* **16**, 467–475. (doi:10.1016/j.tics.2012.07.007)
42. Alexander GE, DeLong MR, Strick PL. 1986 Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* **9**, 357–381. (doi:10.1146/annurev.ne.09.030186.002041)
43. Miller EK, Cohen JD. 2001 An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202. (doi:10.1146/annurev.neuro.24.1.167)
44. Koechlin E, Ody C, Kouneiher F. 2003 The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185. (doi:10.1126/science.1088545)
45. Badre D, Kayser AS, D'Esposito M. 2010 Frontal cortex and the discovery of abstract action rules. *Neuron* **66**, 315–326. (doi:10.1016/j.neuron.2010.03.025)
46. Barbas H, Pandya DN. 1989 Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *J. Comp. Neurol.* **286**, 353–375. (doi:10.1002/cne.902860306)
47. Doya K, Samejima K, Katagiri K, Kawato M. 2002 Multiple model-based reinforcement learning. *Neural Comput.* **14**, 1347–1369. (doi:10.1162/089976602753712972)
48. Cowan N. 2005 Working-memory capacity limits in a theoretical context. In *Human learning and*

memory (eds C Izawa, N Ohta), pp. 155–175. Hillsdale, NJ: Erlbaum.

49. Koechlin E, Basso G, Pietrini P, Panzer S, Grafman J. 1999 The role of the anterior prefrontal cortex in human cognition. *Nature* **399**, 148–151. (doi:10.1038/20178)
50. Sakai K. 2008 Task set and prefrontal cortex. *Annu. Rev. Neurosci.* **31**, 219–245. (doi:10.1146/annurev.neuro.31.060407.125642)
51. Yu A, Dayan P. 2005 Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692. (doi:10.1016/j.neuron.2005.04.026)
52. De Martino B, Fleming SM, Garrett N, Dolan RJ. 2013 Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110. (doi:10.1038/nn.3279)
53. Oberauer K. 2010 Declarative and procedural working memory: common principles, common capacity limits? *Psychol. Belgica* **50**, 277–308. (doi:10.5334/pb-50-3-4-277)
54. Risse S, Oberauer K. 2010 Selection of objects and tasks in working memory. *Q. J. Exp. Psychol.* **63**, 784–804. (doi:10.1080/17470210903147486)
55. Oberauer K. 2002 Access to information in working memory: exploring the focus of attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **28**, 411–421. (doi:10.1037/0278-7393.28.3.411)
56. Donoso M, Collins A, Koechlin E. 2014 Foundations of human reasoning in the prefrontal cortex *Science* **344**, 1481–1486. (doi:10.1126/science.1252254)
57. Genovesio A, Wise S, Passingham R. 2014 Prefrontal-parietal function: from foraging to foresight. *Trends Cogn. Sci.* **18**, 72–81. (doi:10.1016/j.tics.2013.11.007)