Trends in
Neurosciences

CellPress
OPEN ACCESS

Review

Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond

Nir Moneta[1,2,4,*], Shany Grossman[1,4,*], and Nicolas W. Schuck[1,3,*]

The orbitofrontal cortex (OFC) and ventromedial-prefrontal cortex (vmPFC) play a key role in decision-making and encode task states in addition to expected value. We review evidence suggesting a connection between value and state representations and argue that OFC / vmPFC integrate stimulus, context, and outcome information. Comparable encoding principles emerge in late layers of deep reinforcement learning (RL) models, where single nodes exhibit similar forms of mixed-selectivity, which enables flexible readout of relevant variables by downstream neurons. Based on these lines of evidence, we suggest that outcome-maximization leads to complex representational spaces that are insufficiently characterized by linear value signals that have been the focus of most prior research on the topic. Major outstanding questions concern the role of OFC/ vmPFC in learning across tasks, in encoding of task-irrelevant aspects, and the role of hippocampus–PFC interactions.

Computational and neural underpinnings of value-based decision-making

Humans and other mammals are versatile decision-makers, skilled at quickly learning how to achieve their goals in diverse environments. To do so, we learn to anticipate the outcomes of our choices. But while learning outcome expectations is straightforward in simple tasks, optimizing real-world behavior is more complex. It requires generalizing expectations across events and understanding how changing goals affect outcome desirability.

One prominent notion is that the goal of decision-making is to maximize the so-called expected value of a decision [1,2], which is defined as the (time-discounted) sum of all expected future rewards after a choice is made. The basic idea of value maximization goes back centuries to expected utility theory [3], which states that decisions aim to maximize the expected value of a utility function that represents our subjective preference. Expected utility theory has had a marked impact on psychological theory ever since it was observed that rational decision-makers will behave as if they are maximizing expected utility [4], although psychological literature has pointed out many important additional perspectives on what drives human choices [5,6].

In parallel to these discussions, several neuroscientific studies have found that vmPFC and adjacent OFC areas are implicated in value processing in humans, non-human primates, and rodents (e.g., [7,8], for reviews see [9,10] as well as Figure 1A), and interact with the wider corticolimbic dopaminergic reward system [11]. One influential study showed that when monkeys choose between different quantities of flavored juice or water, single neurons in the OFC reflect the animal's subjective value of the different outcomes [7]. The existence of value signals throughout the vmPFC and OFC has since been confirmed in humans, monkeys, and rodents (e.g., [12]; see

Highlights

Orbitofrontal cortex (OFC) and ventromedial-prefrontal cortex (vmPFC) areas encode a rich diversity of variables that intertwine value, task state, and outcome properties, using mixed selectivity.

A core benefit of this representational space is that it can encode contextual variables that are not directly observable but are required for predicting outcomes.

Task state representations emerge in deep reinforcement learning networks alongside value-like signals, offering insights into why the brain multiplexes value with other task-related variables.

The complex activity patterns in OFC/ vmPFC can be interpreted as an internal representation that maps internal states, sensory observations, and past knowledge onto values and choice preferences.

[1] Institute of Psychology, Universität Hamburg, 20146 Hamburg, Germany
[2] Einstein Center for Neurosciences Berlin, Charité Universitätsmedizin Berlin, 10117, Berlin, Germany
[3] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, 14195 Berlin, Germany
[4] These authors contributed equally to this work

*Correspondence:
moneta@mpib-berlin.mpg.de (N. Moneta),
grossman@mpib-berlin.mpg.de
(S. Grossman), and
nicolas.schuck@uni-hamburg.de
(N.W. Schuck).

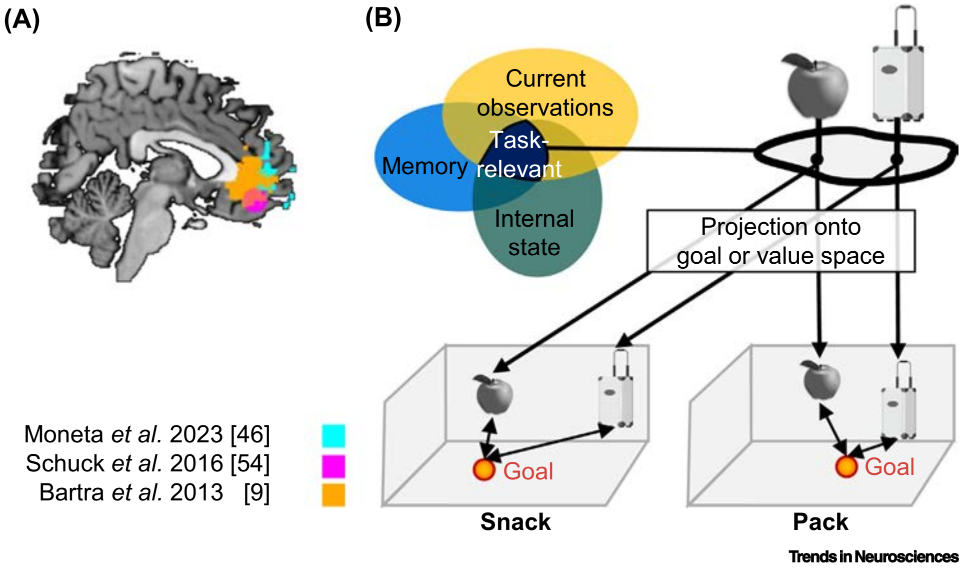


Figure 1. Values and task states in orbitofrontal and ventromedial prefrontal cortex. (A) A midsagittal section of the human brain (MNI template) overlaid with medial prefrontal regions identified as encoding value in a meta analysis of blood oxygen level-dependent (BOLD) fMRI experiments [9] in orange. This region overlaps with the OFC orbitofrontal cortex area reported by Schuck et al. [54] where representations of (partially-observable) states were found (pink), and the medial prefrontal cortex region reported in [46] where value and state representations coexist and interact (cyan). (B) Illustration of how task states influence option values. Task states reflect the combination of sensory and nonsensory variables that are predictive of future outcomes. The computation of task states therefore requires input from several other areas which supply sensory processing, memory function, and access to internal affective and arousal states, amongst others. This information serves to map options onto the values they have for a given goal, thereby allowing the same options to have different values in different contexts. Images of suitcase and apple were adapted from vectorportal (Licensed under CC BY 4.0) and Wikimedia Commons (Licensed under CC BY 1.0), respectively. See [9,46,54].

[8,9] for reviews) and is broadly supported by lesion studies [12–15]. Note that we will use the term expected value hereafter to denote the subjective belief of the subject about the expected outcome of a decision. While in many cases the objective and subjective values align, some experimental paradigms can dissociate the two types of value.

In this review, we argue that the role of OFC/vmPFC goes beyond providing a (subjective) value signal and suggest instead that they have a broader function focused on integrating information in the service of learning to predict outcomes in rich and partially observable environments. We first summarize findings from human, primate, and rodent studies that relate hidden **task state** (see Glossary) representations to these regions and show that value and task state codes are intertwined (Figure 1B). We then show that a similar intertwining occurs in value-maximizing neural networks capable of performing complex tasks. Finally, we argue that these **deep RL models** indicate that value maximizing computations do not necessitate the dominance of value representations as envisaged in neuroscientific research and might serve as a useful model of OFC/vmPFC function that emphasizes the integration of predictive and possibly unobservable task states with expected values.

While our focus is on learning, our conception of OFC/vmPFC function includes a deep interaction with memory processes which can reinstantiate pre-existing value or policy knowledge when needed [16–22]. This process is also crucial when old knowledge needs to be recombined in the service of inference (e.g., [23]) and during continual learning processes that involve ongoing

Glossary

Actor–critic: a policy-based RL algorithm that learns the reward-maximizing probability of choosing among possible actions in a given state of the task (the actor). The model also learns an estimate of the state's value, independent of which action will be chosen, and uses it as a learning signal to optimize the policy (the critic).
Convolutional layer: the building block of convolutional neural networks (CNNs). Each node receives input from a small set of spatially confined nodes (receptive field). With network training, the restricted connectivity leads to nodes acting as filters which detect a specific input feature within their receptive field. Convolutions are typically applied over successive layers, allowing the network to form more complex filters.
Deep Q network (DQN): a value-based deep RL model which receives inputs and maps them to the values of possible actions, with each action being an output node of the network.
Deep reinforcement learning (RL) models: deep neural network models trained with reward signals, instead of supervised teaching signals. This fusion integrates the representation learning abilities of deep learning with the decision-making abilities of RL and allows powerful machine-learning solutions to real life tasks such as autonomous driving.
Fully-connected layer: usually contrasted with usually contrasted with convolutional layer, a fully connected layer is composed of artificial units which are all connected to each other by adjusted weights.
Objective function: the function the model is trained to minimize, usually expressed as the difference between a model prediction, and a target (i.e., what the model should have predicted). For example, a deep RL model can be trained using a Q-loss function, in which the output nodes are trained to match the current and future reward resulting from each action in a given state. In supervised neural networks, the most common objective function is cross entropy.
Recurrent layer: a recurrent layer holds previous observations in its memory and allows them to shape its responses to current inputs. This can be especially useful for solving partially observable tasks, where the estimation of the current state of the world cannot

refinement. We propose that the interaction between OFC/vmPFC and the hippocampus is critical for such a reinstatement.

We also acknowledge that vmPFC and adjacent OFC are anatomically diverse regions with many subdivisions [24]. Although many studies hint at differences between subregions [25], anatomical differences between species and lack of terminological agreement make integration of evidence at a finer anatomical scale difficult. Our focus on the 'OFC/vmPFC' region reflects the most pronounced distinction between medial and lateral areas (e.g., [26]), in line with previous work [27], and broadly corresponds to the 'orbital and medial prefrontal cortex (OMPFC)' region as defined by Öngür and Price [28].

All you need is value?

Inspired by early economic theory (e.g., [4]), some researchers have proposed that value signals reflect a 'common currency' that acts as a stable cardinal desirability scale guiding individuals' decisions [27,29]. According to this conceptualization, one of OMPFC's main functions is to map incommensurable options onto a unidimensional, cardinal scale. Early observations that OFC value signals were independent of sensory features, motor aspects, or other choice options [7,30,31] support this idea, leading to the assumption that OMPFC signals are tailored to generalize across aspects that are represented in other brain areas. This is supported by the finding that vmPFC signals can be decoded across tasks with different goals [17,32–39] and even when cognitive effort [37], or acquisition of knowledge [38] drive valuation or choice.

Other lines of research, however, challenged this idea. Choice preferences, for instance, are affected by irrelevant alternatives, and the range of outcomes [40,41], counter to the predictions of common currency accounts. Contextual information is also important for value encoding, as illustrated in several studies (e.g., [42]; for a review see [43]). Internal states, such as tiredness, modulate choices and also affect values in the brain [44,45], in line with the observation that desirability and neural value signals are goal-dependent; for instance, a hammer is better than a spoon if you want to drive a nail into the wall, while the opposite is true if you want to eat soup [17,33,46–48]. Some evidence also suggests that OMPFC is involved in optimizing other **objective functions**. Decision confidence, for instance, affects OMPFC firing, suggesting that this region might support maximizing confidence [49–52], even when it is orthogonal to expected values [53].

Context matters: how tasks shape choice and neural value signals

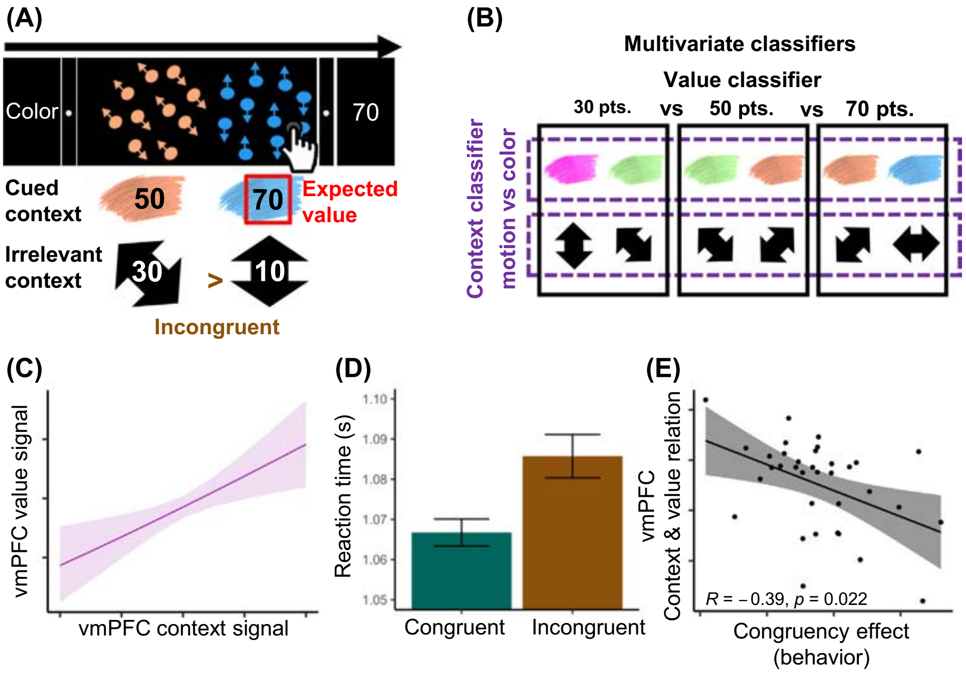
Much work has highlighted context-dependency of decisions, further underlining the aforementioned challenges to common currency ideas (see, e.g., [43,55–58]). Decisions made across contexts, for instance, can systematically violate the principle of value maximization [41,59–61]. In one study, participants were trained to decide between outcomes ranging either from 14 to 50 points or from 14 to 86 points [59]. Asked to pick options across sets, participants chose based on the within-set relative rather than the absolute values, making seemingly irrational decisions. This suggests that values are normalized within each context. Similarly, single cell recordings in macaque OFC found that value signals are normalized by the range of the current context [62,63], in line with human fMRI findings [64], and modeling work [65]. Interestingly, value range adaptation does not seem to appear in OFC during forced choices, suggesting that this form of context-sensitivity is itself context-dependent [66].

One of our recent studies has provided additional insights into the relation between context and value signals in vmPFC [46]. Participants first learned discrete values of four colors and four movement directions while undergoing fMRI. They were then asked to make a choice between two moving and colored stimuli, based only on either the color or motion direction, but not

be fully determined by the current sensory inputs. Recurrent layers are usually contrasted with feed-forward layers which process each observation independently of previous ones.

Task state: collection of observable and non-observable information necessary to predict decision outcomes. The transitions between task states constitute a Markov decision process that allows RL algorithms to solve the reward maximization problem.

both (feature relevance was explicitly cued, and changed every four to seven trials; Figure 2A). Standard analyses confirmed that value was decodable from vmPFC. Nonetheless, using only value-responsive voxels, current task context was also decodable – although contexts were matched in value (Figure 2B). Two keys observations were made: first, value and context were related [i.e., stronger context signals correlated with a stronger value signal within participants (Figure 2C), as well as with the degree to which behavior was influenced by the irrelevant context (Figure 2D)]. Moreover, these two effects were related: a strong connection between vmPFC context and value signals was linked to less influence of 'irrelevant' context on behavior (Figure 2E). Context thus seemed to coexist with and enhance value representations and determined which values influenced behavior. Second, behavior and vmPFC signals were influenced by the irrelevant feature with the highest value, which sometimes was not the chosen option. This implies a hypothetical calculation of the maximal possible value, assuming the alternative context, and possibly another choice (Figure 3A,B). Results hence suggested that vmPFC calculated the



Trends in Neurosciences

Figure 2. Interlinked ventro-medial prefrontal cortex (vmPFC) representations of task state and expected value. (A) Schematic illustration of the experimental paradigm in [46]. After learning to associate rewards with a set of colors and motion directions, participants made choices between two color-motion stimuli. Before each decision, a context cue indicated whether rewards were dependent on color or direction (here: color). The expected value of a trial was the maximum reward of the cued features (here: 70). On congruent trials, choosing the maximally rewarding cued feature also selected the most rewarding uncued feature; on incongruent trials, the reverse was true (see example). Outcome presented after each choice only depended on the features of the cued context. (B) Pattern classifiers were trained on vmPFC data to either distinguish between different values (irrespective of context) or trial context (irrespective of value). The region of interest is indicated in Figure 1A in the main text. (C) Expected value and context could both be decoded from the same vmPFC area, which was defined based on value only (main effect not shown). Moreover, these decoding strengths were related to one another: a stronger context signal (*x-axis*) accompanied a stronger expected value signal (*y-axis*). Shown are mixed effects models testing the association between expected value and context decoding. (D) Participants were slower on incongruent compared with congruent trials (i.e., when the contexts did not agree which decision was best), showing that alternative context influenced behavior. (E) Participants who showed a weaker relationship between context and value representations in vmPFC (*y-axis*, C) also showed a stronger behavioral influence of the irrelevant context (congruency effect, x-axis, D). Plot shows the correlation of the betas from an expected value decoding model (y-axis) with the congruency effect in reaction times (x-axis). Panels modified from [46].

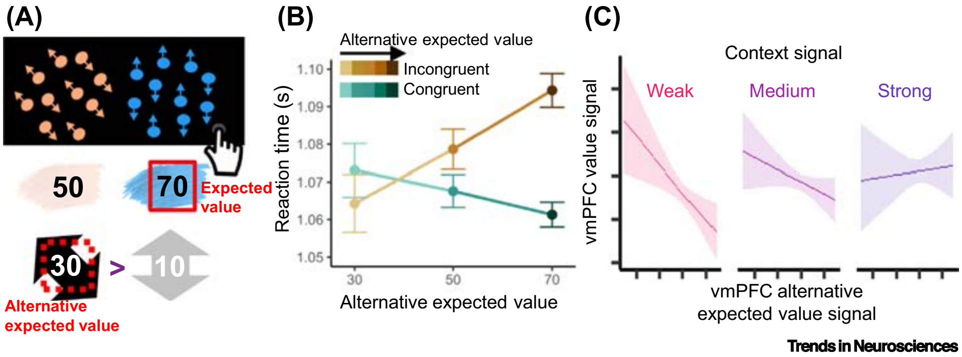


Figure 3. Context encoding modulates irrelevant value signals. (A) Schematic illustration of the experimental paradigm in [46]. Participants made choices between two color-motion stimuli, cued to focus only on color or only on motion. Reward was only predicted by the cued context (here, color). In the trial presented, the best choice according to the cued dimension is on the right (denoted expected value), while the maximum rewarding feature of the cued-to-ignore context is on the left, making this trial incongruent. Thus, the 'alternative' expected value reflects the maximum number of points that would have been obtained in the alternative context (i.e., if the irrelevant features were the relevant ones). (B) Alternative expected value influenced behavior, but only in relation to the congruency of contexts. Larger irrelevant values led to faster reaction times in congruent trials (green), and slower reaction times on incongruent trials (brown). Note that on incongruent trials, the irrelevant expected value reflected a different, hypothetical choice (cf. panel A). (C) Relationship between neural signals for expected value (*y-axis*) and the strength of irrelevant expected value (*x-axis*), separately for trials in which the context signal was weak, medium, or strong. The neural representation of both expected values (stemming from the relevant and irrelevant contexts) was negatively related. However, this negative relation was modulated by the context signal. When the context signal was strong, the influence of the irrelevant values on the ventro-medial prefrontal cortex (vmPFC) signal was reduced, akin to an arbitrating effect between competing value signals. Shown relationships reflects mixed effects model testing the association between expected value decoding and alternative expected value decoding.

values of each context, which then competed for representation. Strikingly, context signals modulated which value signal dominated vmPFC – the true value or the aforementioned hypothetical value (Figure 3C). Hence, task-context signals, not options or actions, organized value representations and choices (Figure 3 and Box 1), providing a way for vmPFC value signals to reflect possible future goals in addition to ongoing tasks.

Box 1. Representation and compression of task-irrelevant values

A major role of state representations is to define which information is needed to predict outcomes in a given context. State representations have been linked to OMPFC [54,74], where lesions hinder the ability to ignore irrelevant choice options [142]. Moreover, OMPFC activity compresses inputs to focus on goal-relevant information [82,104,143] and guides hippocampus in forming reward-predictive relational maps [93]. Compression also occurs in deep neural networks, although contingent on factors such as activation functions [144] or weight initialization [145]. This raises the question how complete such compression is, that is, if OMPFC still maintains some representation of (i) irrelevant task/stimulus information, and (ii) irrelevant values.

In a study addressing the first question, participants were initially instructed to focus on one of two stimulus features, but later, unbeknownst to participants, the previously irrelevant feature suddenly became task-relevant [139]. Some participants did notice the changed relevance – even when this was never needed to complete the task [146] – and MRI results showed irrelevant information processing arose in mPFC before participants abruptly changed choices to tap into changed relevance (note that the region found in this study was more dorsal than commonly seen in value studies). Neural network simulations of the same task demonstrate that regularized gating can lead to preserved latent knowledge of irrelevant aspects which can be accessed rapidly if needed and leads to similarly abrupt and spontaneous behavioral switches as those observed in humans [147,148]. Broadly in line with this idea, some studies have shown that task-irrelevant features can be decoded from the frontal eye fields in monkeys and motor cortex in humans [149,150].

If some representation of task-irrelevant information is maintained, what happens to task-irrelevant If some representation of task-irrelevant information is maintained, what happens to task-irrelevant values values ? Many fMRI studies reported task-relevant values within vmPFC, but no univariate evidence of task-irrelevant values has been found ([17,33,78,79]; but see for [16–19,151] for task-independent value-like signals). However, using multivariate methods, recent work from our group showed that such task-irrelevant values do exist in vmPFC [46], interact with other value and non-value representations, and influence behavior (Figures 2 and 3). This raises the possibility that multivariate fMRI methods are better suited to uncover compressed representations.

That vmPFC representations of contexts and values interact with each other, is broadly in line with evidence for context signals alongside values in OMPFC [36,67–69], but also extends previous work by demonstrating an intricate interplay of value and context signals.

From context to task states: how cognitive maps influence values

The context-sensitivity of value signals in the brain might not be surprising given that adaptive behavior needs to reflect how goals and contexts influence outcome desirability. But how exactly should 'context' be defined? One perspective, albeit not the only one, comes from RL theory [2], which formalizes how agents can learn reward maximizing behavior from trial-by-trial feedback. The simplest RL algorithms receive handcrafted information about the current 'state', or context, of the environment, which does not have to be directly observable but can for instance be defined by past events or internal needs. If RL models activate the wrong state they will also retrieve the wrong value, which means that reward learning is always contingent on current state knowledge (Figure 1B). While this perspective agrees with other theories that values are abstract in nature and enable comparison of incommensurable options, it suggests that relevant task details might exist in the same region – a level of specificity that has been de-emphasized in particular by common currency approaches. The aforementioned findings of context signals that reside alongside value in OMPFC [36,54,67–69] and their close connection to value signals in the same area [46] support this perspective.

What are states, specifically? First, RL accounts emphasize that states must (exclusively) reflect information that is needed to predict future reward. This can be sensory information (whether the sun is shining or not), but it can also be something that cannot be observed directly, such as how much time has passed. The second major aspect is that, in some RL models, states are part of a cognitive map that specifies transitions between them. This emphasis on predicting future states is core to RL approaches that provide additional flexibility (e.g., model-based RL [70]; successor representations [71] or replay approaches [72], see later). In sum, the RL perspective therefore emphasizes the role of reward and state predictiveness as defining features of context. Both aspects are in line with research on state representations in OMPFC [54,73–75] as well as on OFC role in generalization and inference [67,76,77]. More broadly, it can also explain the dominance of goal-aligned value signals in OMPFC in cases where values depend on goal context [17,33,46,78,79] (see later).

Adopting this perspective, one study showed that OFC lesions in rats affect reward-related dopamine firing in line with predictions from an account that assumes OFC is needed to signal, latent task states that are independent from sensory input (known as 'partially observable') [74]. An fMRI study in humans also found that partially observable states can be decoded from medial OFC [54], in line with a number of other studies [46–48,69,75,80–82]. Hence, OMPFC may infer latent states that are needed to retrieve context-sensitive values, which is crucial when the same choices lead to different outcomes given partially observable states. Some studies suggest that OFC represents task structures and rules even without any explicit value [54,83,84], akin to 'schemas' that also seem to reside in OMPFC [85,86]. OFC representations also appear similar when the same task is done with and without rewards, presumably reflecting stimulus–stimulus associations [87], akin to latent learning ideas [88] that emphasize how stimulus–stimulus learning done in the absence of rewards can be used for later reward tasks. This suggests that OMPFC might serve more broadly as a cognitive map that guides decisions [89–91], a function that is likely to occur in close connection with the hippocampus (see later for further discussion and cf. [92–95]).

In sum, a state representation perspective envisages a dynamic process in which choice options can be flexibly projected onto different expected values, depending on goals and past history that influences how desirable an option is. States also support efficient learning by forming a cognitive map that facilitates generalization and planning in information-rich environments with complex temporal structure (see e.g., [73,91,96]).

While the evidence discussed earlier generally supports this perspective, several questions remain. One notable deviation is that OMPFC systematically represents task-irrelevant information and even 'irrelevant' values – a finding that we discuss in Box 1. A second issue concerns cases in which non-value quantities are optimized, such as distance to a non-value goal (see later), where the predictiveness of OMPFC signals will not refer to future rewards, but closeness to goal. Finally, we acknowledge that while some have presented direct computational or empirical evidence that context and cognitive maps can be understood as states that arise in RL machinery (e.g., [71,97]), further evidence is needed.

A complex and versatile code in OMPFC

The evidence reviewed so far suggests that values are part of a complex activation manifold with multiple dimensions related to choice values, (partially observable) task states, and alternative values [46,54,62,63,74]. Electrophysiological studies support this idea and indicate diverse information encoding in OFC [69,98,99]. For instance, neurons in OMPFC encode summary statistics of the current task such as previous offers and outcomes, or the location of the currently attended offer [100]. Recording studies have also shown that the same neurons in OFC often encode multiple variables at once, a phenomenon known as mixed selectivity [101]. One study recorded neurons from monkeys performing a choice between options characterized by the flavor and probability of a juice reward. The authors reported that most neurons in OFC showed mixed selectivity for probability and flavor [102]. Another study showed that the same neurons in macaque OFC can represent both spatial and reward information, even when those are unrelated [103]. Although different variables can be encoded by the same neurons [104] or voxels [46], merely representing different variables does not mean that they are integrated. Perhaps the most direct evidence for such an integration comes from [46] (see earlier), where context signal strength covaried with value signal strength and behavioral markers or context adaptive behavior.

This line of evidence raises a major conundrum: how can the findings that OMPFC activity multiplexes many task variables with reward expectations be reconciled with the reports of generalizable, content-independent value representations discussed earlier? One explanation could be that, on a population level, neurons with mixed selectivity can still form a high dimensional representation in which mostly orthogonal planes reflect different variables [105]. This means that downstream neurons can easily ready out independent codes for each variable. Indeed, in [102] the subspaces of population activity within OFC reflected probability and flavor and were minimally dependent (i.e., nearly orthogonal). We note that mixed selectivity is not unique to OMPFC and prevalent throughout the frontal cortex [101,106]. This suggests that mixed selectivity has a very broad function in high-order and flexible cognition that goes beyond the specific computations in OMPFC. Another, not exclusive, possibility is that goal-independent representations emerge during a late computational stage when state-dependent values transform into specific action-selection signals. In support of this idea, it has been shown that while the same OFC neuron population can be involved in evaluation and selection during value-guided choice, activations during these different phases lie on almost orthogonal subspaces [107]. Others found expected values in OMPFC arise only when the task requires a selection ([33], but see [16]), without encoding motor signals [7,46,58,108].

A neural network perspective on value and state representations

As reviewed earlier, OMPFC does not appear to be exclusively committed to signaling only value or only task states. But can an exclusive focus on univariate codes for either value or task states even be expected from a complex computational system like the brain? One avenue for addressing this question is to study deep RL models, which reflect an integration of RL with deep neural networks [109]. In addition to being powerful AI tools that can master games [110,111], or control self-driving cars [112], deep RL models can be useful for neuroscientific research of learning and decision-making (Box 2).

Deep RL models' advantage over classic RL is their ability to master reward learning in complex environments by learning task-specific representations in their late layers, positioned close to the output. Classic RL models directly receive information about a small number of hand-crafted and discrete states, such as their position in an artificially discretized spatial environment. Deep RL models, by contrast, can learn directly from a high-dimensional, continuous, and noisy sensory description (e.g., perceived distance to the wall) by relying on their representation learning power to form task-appropriate lower dimensional representations of the input [113] (Figure 4). Critically, these emergent representations arise without explicit guidance other than the network's objective function, which for deep RL is typically the optimal reward (e.g., [110]) or policy (e.g., [100,114,115]), and they often end up having many features of the (partially-observable) task states we reviewed earlier. A reward-focused objective function therefore does not only lead to value representations, but also to task-appropriate abstractions of the sensory input, as we discuss later.

Because of their densely connected multilayer architecture, deep RL models also learn differently. Standard models only update the currently activated state when receiving input. But in deep RL models input and output are connected via many intermediate hidden layers that often feature mixed-selectivity [116–119], similar to what has been observed in the OMPFC (see earlier). A single weight-update will therefore affect many representations, and learning is never confined to just one state (Figure 4B).

Box 2. Using deep RL models for the neuroscientific study of decision-making and learning

The core idea of deep RL models is to train a deep neural network through trial-and-error reward feedback, rather than through supervised training. These models usually receive sensory observations as inputs, such as image pixels, and are trained to output expected values and/or actions that maximize reward. Most deep RL models process visual inputs in early **convolutional layers**, on top of which fully-connected layers are stacked. A popular type are deep Q networks (DQNs), which approximate the expected values of a set of discrete actions, given the input. Important additions to this standard architecture are recurrent layers that provide the network with memory and replay buffers that allow offline sampling [109,138].

Deep RL models are regarded as a useful tool in neuroscience because they share some basic properties with the brain. They process information through layers of connected and distributed nodes in a stage-like fashion and learn by adjusting the connection strength between nodes as a function of feedback. These broad principles are reminiscent of the distributed information processing and synaptic plasticity found in real neurons. Although these similarities are relatively superficial – substantial differences exist, for instance, in how synaptic weight updates are propagated throughout the networks – the main promise of deep RL networks is to offer a useful level of abstraction for studying algorithmic aspects of cognition. Because deep RL models can perform complex cognitive tasks on par with humans [110,124], they seem to retain at least some of the necessary ingredients for complex cognitive skills. A burgeoning field now uses deep RL as models for behavior [152]. Some work has made progress by deriving analytical solutions of learning dynamics in simplified neural networks which yield explanations for observed learning trajectories [153]. Others have used them to derive testable neurobiological predictions about context-dependent learning [145] or to provide explanations about why certain computational ingredients are essential for achieving human-like learning [64,154]. Finally, deep RL models are useful because they allow studying the interaction of learning algorithms, behavior, and representations, providing, for instance, ideas about which representations can be expected in value maximizing networks. One example that showcases this strength comes from the area of distributional RL models [155], which suggests benefits of computing many diversely tuned reward prediction errors (RPEs), rather than only the single RPE assumed in standard RL. While the single RPE signal has famously been found in the firing rate of dopamine neurons [156], the notion of distributional RL in midbrain circuits has recently gained empirical support [119,157]. In a similar vein, deep RL models might help refine understanding of OMPFC codes in the brain.

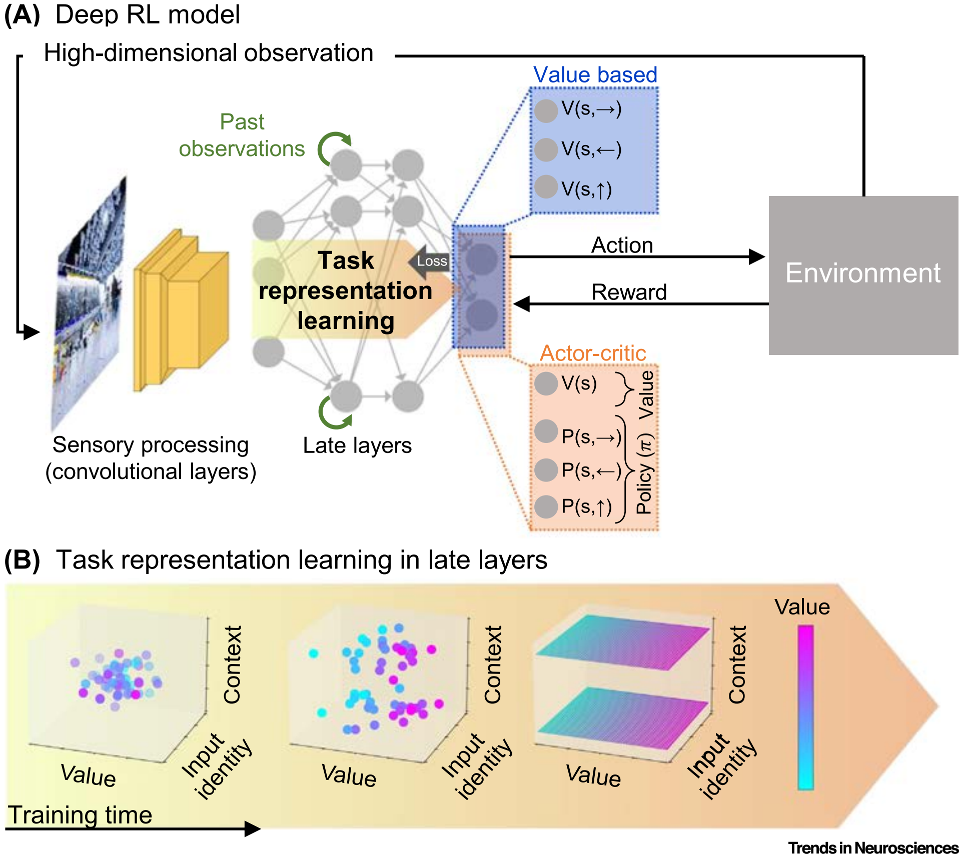


Figure 4. Principles of deep reinforcement learning (RL) models and emergent representations in late hidden layers (A) General scheme of a deep RL model. High-dimensional inputs (e.g., thousands of pixels) are first processed through stacked convolutional layers (akin to sensory processing), usually followed by non-convolutional fully-connected layers ('late layers', marked here by the yellow-orange arrow). Adding recurrency to the network (green arrows) allows to incorporate past events with present observations for representing partially-observable states. In the case of purely value-based models, such as deep Q networks, the output nodes are trained to approximate the expected action values; for more policy oriented models, output nodes are trained to reflect action probabilities, sometimes in addition to value estimates (such as in actor–critic). Once an action is chosen and rewarded, the error is used to update the weights of the model, incrementally forming hidden representations. While the learning process is focused on achieving maximally accurate value or action probability estimates, it also shapes the representations in the late layers such as to form distributed representations of task and value relevant variables that represent a compact world model, or 'cognitive map', of the task. (B) A schematic exemplifying the formation of hidden task representations in late layers of the deep RL network, while multiplexing value with non-value task variables. The axes reflect reduced dimensions of the population code (e.g., through principal component analysis or other dimensionality reduction methods). Note that context is not explicitly signaled in the input and the network needs to infer contexts, often based on observations that go beyond those currently observed (as is the case in partially-observed tasks).

Value and state representations in deep RL models

Using the principles outlined earlier, deep RL models solve complex tasks by learning to extract multiple layers of representations, with an increasing level of abstraction [120]. A major question is what characterizes late layer representations that are only a few computations away from the decision output, and whether their features correspond to what we know about OMPFC. One notable paper has shown that a recurrent deep RL model can capture several core aspects of OFC function and might reconcile value and state accounts on this brain area [121]. Other evidence from neural network studies also suggests late layer representations are not merely sorted by the value of the input they correspond to. In one study, [110] a **deep Q network**

(DQN) (a form of deep RL model) was trained to play various Atari games. Visually investigating the DQN's last layer representations showed that the different input frames were not uniformly sorted by the value that the network predicted for them. Rather, value led to some clustering, but other factors such as perceptual or strategic similarity were reflected too. A subsequent study [122] found that the geometry of representations in late layers of the same DQN (i.e., their pairwise similarities) correlated with a hand-crafted geometry which retains abstract information about input features (e.g., the ball position, or the position of the two paddles in the game pong). Such abstract signals are reminiscent of the task states discussed earlier. Counter to our argument, however, in [122], no link was found between these representations and human participants' OMPFC activity.

Other work has used encoding models to ask whether hidden units reflect human-generated concepts. One study [123] demonstrated this approach on a deep RL model for chess, AlphaZero [124]. The authors found that late hidden layers come to represent many concepts other than the expected value of the current board, such as whether the player is in check, or whether the opponent can capture the queen. The selectivity profile of single units of deep RL models portrays a similar picture. In a deep RL model trained to solve a spatial reward task, a recent study [125] found that good performance was related to the emergence of value selective units. But these units made up only 10% to 50% of the population and units not related to value were also highly correlated with the performance of the model (although this analysis was performed on units taken from all four model layers, so its result is not exclusive to the late layers).

Similar observations have been made for different deep RL architectures. Hidden units in a decision network of an actor–critic-based recurrent RL model (Box 2) show mixed selectivity to combinations of task conditions, such as context and stimulus coherence levels [116]. Other work on recurrent RL models has shown that hidden representations capture task structure by retaining information about recent choices and rewards [118,126,127]. Another line of work has demonstrated the importance of non-value representations more broadly by showing that adding other constraints on hidden representations than a reward maximization objective helps performance [128,129]. Unsupervised pretraining of neural networks, for instance, can speed up later training with a specific objective function [130]. In addition, maximizing mutual information between hidden representations of inputs that are adjacent in time and space can enable better abstraction and generalization in Atari games [131]. Further, a deep RL architecture can benefit from being endowed with grid-like representations prior to learning [132] or self-supervised learning objectives [133].

In sum, we argue that the late layers of deep RL networks offer a useful model to understand the computational role of OMPFC. This role consists in using (reward) feedback signals to shape a mixed selectivity code in a way that emphasizes outcome predictive state and value representations. This process depends critically on input from many other regions, which, for instance, provide appropriately processed sensory information, or access to working memory. Perhaps the biggest challenge to this idea is that studies that directly compare late layers in deep RL models and OMPFC signals are largely lacking and available evidence is inconclusive. Future experiments should test the extent to which deep RL models truly align with observed OMPFC signals.

Beyond standard deep RL: flexibility through long-term memory, meta learning, and model-based RL

If OMPFC signals reflect computations akin to late layers in deep RL models, how can this account for value or liking signals in OMPFC that occur when no learning or action are needed? A potential explanation is that once value information has been established (e.g., value of known food items), the information can be reinstantiated in the network, for instance when conditioned stimuli are presented. This reinstatement is also critical for inference or when past goals are revisited and may even occur spontaneously in the absence of a choice task.

The importance of flexible access to long-term memory has often been overlooked because most laboratory tasks capture 'isolated' learning processes that start from a blank slate and are completed after a few hours of experience. Yet, the perhaps most remarkable aspect of animal and human learning is the flexibility with which subjects apply previously gained insights to new problems [134] and learn over long time spans to extract commonalities between learning problems [135]. A notable recent advance that captures this idea, and can make deep RL models more flexible, is deep 'meta-RL'. Deep meta-RL uses a meta-learning approach in which slow weight changes enable fast learning in the activity dynamics of a recurrent network [136]. The resulting models capture instances of accelerating learning over a set of new but related problems [127]. Notably, recent findings have shown that plasticity within OFC is necessary for such a process [126]. Hence, the combination of access to previously established knowledge with the aforementioned learning powers could give OMPFC a unique power to meta-learn and integrate fast with slow learning processes. One aspect of neural processing of potential relevance for this notion is the interaction between OMPFC and hippocampus, given the important role of the hippocampus in long-term memory and memory reactivation, the functional similarities between both regions, and their close connectivity [28]. It should be noted that this idea suggests a more complex deep RL architecture with separate long-term value storage systems that interact with OMPFC.

A second approach within RL frameworks to support flexibility when adapting to new problems is model-based RL, which learns a model of state transitions separately from values. The combination of values and a state transition model can then be used to make on-the-fly value calculations. Yet, most deep RL models we discussed so far are in fact model-free (i.e., they do not incorporate structural knowledge) and consequently tend to be inflexible. When it comes to the brain, one possibility is that transition knowledge is stored outside of the OMPFC, but can still influence OMPFC computations via offline updating [137]. In line with this idea, previous work has found that replay in the hippocampus – a putative mechanism used by the brain to sample from a model of the task during rest [138] – is linked to state representations in the OFC [72], suggesting a role of hippocampal–OFC interactions in the service of flexibility.

A final consideration concerns the availability of task-irrelevant signals in deep RL models that is in line with findings about irrelevant signals in the OMPFC discussed earlier [18,19,46,139]. An intriguing open question is whether such irrelevant signals are intentionally retained to accommodate for a dynamic environment with constantly changing contexts ('a feature'), or whether the computational machinery is limited in suppressing them fully ('a bug'). Further studying such cross-task signals in deep RL models trained on several tasks might help elucidate the origins of their neural counterparts.

Concluding remarks and future perspectives

We provided an overview of information encoded in OMPFC during decision-making tasks. OMPFC representations are multifaceted, shaped not only by immediate and expected rewards, but also by sensory and non-sensory information required for optimizing behavior in current tasks. As discussed in previous sections, such representational richness aligns with the concept of task states in RL and with late hidden layer activations that arise in deep RL models that learn to perform complex tasks. Ultimately, this suggests that value-oriented computations do not necessarily lead to simple representations of expected value in the form of a universal currency for decision-making. Instead, we propose a perspective in which the OMPFC provides an integration of value and task states in the service of decision-making in complex environments. We also highlighted the important observation from neurophysiological as well as simulation work that single

Outstanding questions

Do late layers of deep RL models offer a model for representations in OMPFC? If so, which network architectures and layers therein best match OMPFC regions?

Can recurrent neural network architectures reveal previously unidentified links between deep RL models and OMPFC, given their role in partially-observable tasks and meta learning?

What influences encoding of non-value signals and information irrelevant for the current task in OMPFC? Can this be linked to objective functions, weight initialization, and activation functions in neural networks?

What controls the amount of compression and (dis)entanglement in OMPFC representations?

Learning is a dynamic process that evolves over time. Is OMPFC's role in decision-making most prominent during early learning stages?

How do the representation learning dynamics of deep RL on a trial-by-trial level, and across episodes, compare with those in OMPFC?

Does OMPFC guide value-free decision-making processes in the brain? How can one disentangle value-based versus value-free learning and choice?

How do ventromedial and orbital areas interact with the hippocampus during offline replay and during on-task periods to guide decision-making?

neurons are characterized by mixed selectivity to linear and nonlinear mixtures of value, outcome, task state, and other variables. Notably, while information is mixed on the single neuron level, it is still possible for different variables to be read out independently on the population level (example visualized in Figure 4B). This implies that complex neural codes which feature information integration on the single neuron level do not contradict the existence of more abstract, independent representations on the population level [106]

While deep RL can offer useful insights for OMPFC function, we believe a number of important aspects need to be considered. Of particular relevance is the need to complement on-task learning powers of standard models with access to long-term memory in a way that enables learning across tasks over larger horizons. Some promising first results have indicated a link between OMPFC and meta-learning and memory replay, but more work will be needed, in particular concerning the role of hippocampus–OMPFC interactions in this regard. We also argue that reinstatement of established (value) knowledge could explain the documented role OMPFC plays for flexible generalization, as well as in tasks that neither require learning nor choices. Another important observation that requires deeper investigation is that value signals can reflect not only current but also future or hypothetical tasks (see [46]), suggesting OMPFC decision-making function reflects not only past tasks, but also future ones.

While neural network-based computational models might inspire new concepts and predictions concerning representations in OMPFC, another remaining challenge is the lack of a clear correspondence between network components or computations and specific brain regions. Most observations about similarities of deep RL models and OMPFC remain qualitative, and a previous study [122] failed to find any direct relation between the model representations and fMRI activity in OMPFC. Additional in-depth investigations are therefore critical (see Outstanding questions).

Finally, we believe that it is time to reconceptualize value as a multidimensional signal that tracks distance to the current task-goal, rather than accumulated reward [17,43,140]. This approach could open the door for frameworks that integrate goal and value signals [141], integrate confidence into the decision [49–52], and even ones which assume no explicit computation of value at all [55]. Together, we believe these shifts in focus will help gain better understanding of the full complexity of OFC/vmPFC function.

Acknowledgments
We thank Chris Summerfield, Ondrej Zika, and Alexa Ruel for valuable comments on this manuscript. N.W.S. is funded by a Starting Grant from the European Union (ERC-2019-17 StG REPLAY-852669) and the Federal Ministry of Education and Research (BMBF) and the Free and Hanseatic City of Hamburg under the Excellence Strategy of the Federal Government and the Länder. S.G is funded by the Zuckerman-CHE female postdoctoral program.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance language and readability in a few instances. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Silver, D. *et al.* (2021) Reward is enough. *Artif. Intell.* 299, 103535
2. Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, MIT Press
3. Peasgood, T. (2014) *Expected Utility Theory*, Springer, pp. 2092–2096
4. Samuelson, P.A. (1947) Some implications of “ linearity”. *Rev. Econ. Stud.* 15, 88–90

5. Gigerenzer, G. and Gaissmaier, W. (2011) Heuristic decision making. *Annu. Rev. Psychol.* 62, 451–482
6. Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291
7. Padoa-Schioppa, C. and Assad, J.A. (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226
8. O'Doherty, J. *et al.* (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* 4, 95–102
9. Bartra, O. *et al.* (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* 76, 412–427
10. Clithero, J.A. and Rangel, A. (2014) Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* 9, 1289–1302
11. Averbeck, B. and O'Doherty, J.P. (2022) Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology* 47, 147–162
12. Ballesta, S. *et al.* (2020) Values encoded in orbitofrontal cortex are causally related to economic choices. *Nature* 588, 450–453
13. Fellows, L.K. (2007) The role of orbitofrontal cortex in decision making. *Ann. N. Y. Acad. Sci.* 1121, 421–430
14. Hogeveen, J. *et al.* (2017) Impaired valuation leads to increased apathy following ventromedial prefrontal cortex damage. *Cereb. Cortex* 27, 1401–1408
15. Vaidya, A.R. and Fellows, L.K. (2020) Under construction: ventral and lateral frontal lobe contributions to value-based decision-making and learning. *F1000Res* 9, F1000 Faculty Rev-158
16. Lebreton, M. *et al.* (2009) An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* 64, 431–439
17. Frömer, R. *et al.* (2019) Goal congruency dominates reward value in accounting for behavioral and neural correlates of value-based decision-making. *Nat. Commun.* 10, 4926
18. Abitbol, R. *et al.* (2015) Neural mechanisms underlying contextual dependency of subjective values: converging evidence from monkeys and humans. *J. Neurosci.* 35, 2308–2320
19. Harvey, A.H. *et al.* (2010) Monetary favors and their influence on neural responses and revealed preference. *J. Neurosci.* 30, 9597–9602
20. Suzuki, S. *et al.* (2017) Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nat. Neurosci.* 20, 1780–1786
21. Lopez-Persem, A. *et al.* (2020) Four core properties of the human brain valuation system demonstrated in intracranial signals. *Nat. Neurosci.* 23, 664–675
22. Plassmann, H. *et al.* (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.* 27, 9984–9988
23. Barron, H.C. *et al.* (2020) Neuronal computation underlying inferential reasoning in humans and mice. *Cell* 183, 228–243
24. Cavada, C. *et al.* (2000) The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cereb. Cortex* 10, 220–242
25. Wang, M.Z. *et al.* (2022) A structural and functional subdivision in central orbitofrontal cortex. *Nat. Commun.* 13, 3623
26. Izquierdo, Alicia (2017) Functional heterogeneity within rat orbitofrontal cortex in reward learning and decision making. *J. Neurosci.* 37, 10529–10540
27. Levy, D.J. and Glimcher, P.W. (2012) The root of all value: a neural common currency for choice. *Curr. Opin. Neurobiol.* 22, 1027–1038
28. Öngür, D. and Price, J.L. (2000) The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cereb. Cortex* 10, 206–219
29. Fehr, E. and Rangel, A. (2011) Neuroeconomic foundations of economic choice –recent advances. *J. Econ. Perspect.* 25, 3–30
30. Padoa-Schioppa, C. and Assad, J.A. (2008) The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat. Neurosci.* 11, 95–102
31. Tremblay, L. and Schultz, W. (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* 398, 704–708
32. Gross, J. *et al.* (2014) Value signals in the prefrontal cortex predict individual preferences across reward categories. *J. Neurosci.* 34, 7580–7586
33. Castegnetti, G. *et al.* (2021) How usefulness shapes neural representations during goal-directed behavior. *Sci. Adv.* 7, eabd5363
34. McNamee, D. *et al.* (2013) Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nat. Neurosci.* 16, 479–485
35. Yao, Y.-W. *et al.* (2023) The dorsomedial prefrontal cortex represents subjective value across effort-based and risky decision-making. *NeuroImage* 279, 120326
36. Zhang, Z. *et al.* (2017) Distributed neural representation of saliency controlled value and category during anticipation of rewards and punishments. *Nat. Commun.* 8, 1907
37. Westbrook, A. *et al.* (2019) The subjective value of cognitive effort is encoded by a domain - general valuation network. *J. Neurosci.* 39, 3934–3947
38. Kobayashi, K. and Hsu, M. (2019) Common neural code for reward and information value. *Proc. Natl. Acad. Sci. U. S. A.* 116, 13061–13066
39. Howard, J.D. *et al.* (2015) Identity-specific coding of future rewards in the human orbitofrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* 112, 5195–5200
40. Vlaev, I. *et al.* (2011) Does the brain calculate value? *Trends Cogn. Sci.* 15, 546–554
41. Bavard, S. *et al.* (2018) Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nat. Commun.* 9, 4503
42. Winston, J.S. *et al.* (2014) Relative valuation of pain in human orbitofrontal cortex. *J. Neurosci.* 34, 14526–14535
43. Juechems, K. and Summerfield, C. (2019) Where does value come from? *Trends Cogn. Sci.* 23, 836–850
44. Pastor-Bernier, A. *et al.* (2021) Reward-specific satiety affects subjective value signals in orbitofrontal cortex during multicomponent economic choice. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2022650118
45. Yoshimoto, T. *et al.* (2022) Coexistence of sensory qualities and value representations in human orbitofrontal cortex. *Neurosci. Res.* 180, 48–57
46. Moneta, N. *et al.* (2023) Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. *Nat. Commun.* 14, 3156
47. Zhou, J. *et al.* (2019) Rat orbitofrontalensemble activity contains multiplexed but dissociable representations of value and task structure in an odor sequence task. *Curr. Biol.* 29, 897–907.e3
48. Wimmer, G.E. and Büchel, C. (2019) Learning of distant state predictions by the orbitofrontal cortex in humans. *Nat. Commun.* 10, 2554
49. De Martino, B. *et al.* (2013) Confidence in value-based choice. *Nat. Neurosci.* 16, 105–110
50. Gherman, S. and Philiastides, M.G. (2018) Human vmPFC encodes early signatures of confidence in perceptual decisions. *eLife* 7, e38293
51. Lebreton, M. *et al.* (2015) Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* 18, 1159–1167
52. Barron, H.C. *et al.* (2015) Reassessing vmPFC: full of confidence? *Nat. Neurosci.* 18, 1064–1066
53. Shapiro, A.D. and Grafton, S.T. (2020) Subjective value then confidence in human ventromedial prefrontal cortex. *PLoS One* 15, e0225617
54. Schuck, N.W. *et al.* (2016) Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* 91, 1402–1412
55. Hayden, B.Y. and Niv, Y. (2021) The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* 135, 192–201
56. Miller, K.J. *et al.* (2019) Habits without values. *Psychol. Rev.* 126, 292–311
57. Palminteri, S. and Lebreton, M. (2021) Context-dependent outcome encoding in human reinforcement learning. *Curr. Opin. Behav. Sci.* 41, 144–151
58. Knudsen, E.B. and Wallis, J.D. (2022) Taking stock of value in the orbitofrontal cortex. *Nat. Rev. Neurosci.* 23, 428–438
59. Bavard, S. and Palminteri, S. (2023) The functional form of value normalization in human reinforcement learning. *eLife* 12, e83891

60. Molinaro, G. and Collins, A.G.E. (2023) Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS Biol.* 21, e3002201
61. Palminteri, S. *et al.* (2015) Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* 6, 8096
62. Conen, K.E. and Padoa-Schioppa, C. (2019) Partial adaptation to the value range in the macaque orbitofrontal cortex. *J. Neurosci.* 39, 3498–3513
63. Padoa-Schioppa, C. (2009) Range-adapting representation of economic value in the orbitofrontal cortex. *J. Neurosci.* 29, 14004–14014
64. Nelli, S. *et al.* (2023) Neural knowledge assembly in humans and neural networks. *Neuron* 111, 1504–1516.e9
65. Zimmermann, J. *et al.* (2018) Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nat. Commun.* 9, 3206
66. Yamada, H. *et al.* (2018) Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nat. Commun.* 9, 162
67. Baram, A.B. *et al.* (2021) Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron* 109, 713–723.e7
68. Cromwell, H.C. *et al.* (2018) Neural encoding of choice during a delayed response task in primate striatum and orbitofrontal cortex. *Exp. Brain Res.* 236, 1679–1688
69. Farovik, A. *et al.* (2015) Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. *J. Neurosci.* 35, 8333–8344
70. Sutton, R.S. (1991) Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull.* 2, 160–163
71. Stachenfeld, K.L. *et al.* (2017) The hippocampus as a predictive map. *Nat. Neurosci.* 20, 1643–1653
72. Schuck, N.W. and Niv, Y. (2019) Sequential replay of nonspatial task states in the human hippocampus. *Science* 364, eaaw5181
73. Niv, Y. (2019) Learning task-state representations. *Nat. Neurosci.* 22, 1544–1553
74. Wilson, R.C. *et al.* (2014) Orbitofrontal cortex as a cognitive map of task space. *Neuron* 81, 267–279
75. Bradfield, L.A. and Hart, G. (2020) Rodent medial and lateral orbitofrontal cortices represent unique components of cognitive maps of task space. *Neurosci. Biobehav. Rev.* 108, 287–294
76. Shi, W. *et al.* (2023) The orbitofrontal cortex: a goal-directed cognitive map framework for social and non-social behaviors. *Neurobiol. Learn. Mem.* 203, 107793
77. Boorman, E.D. *et al.* (2021) The orbital frontal cortex, task structure, and inference. *Behav. Neurosci.* 135, 291–300
78. Grueschow, M. *et al.* (2015) Automatic versus choice - dependent value representations in the human brain. *Neuron* 85, 874–885
79. Hare, T.A. *et al.* (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324, 646–648
80. Chan, S.C.Y. *et al.* (2016) A probability distribution over latent causes, in the orbitofrontal cortex. *J. Neurosci.* 36, 7817–7828
81. Costa, K.M. *et al.* (2023) The role of the lateral orbitofrontal cortex in creating cognitive maps. *Nat. Neurosci.* 26, 107–115
82. Muhle-Karbe, P.S. *et al.* (2023) Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron* 111, 3885–3899.e6
83. Lipton, P.A. *et al.* (1999) Crossmodal associative memory representations in rodent orbitofrontal cortex. *Neuron* 22, 349–359
84. Zhou, J. *et al.* (2021) Evolving schema representations in orbitofrontal ensembles during learning. *Nature* 590, 606–611
85. O. Bein and Y. Niv. Schemas, reinforcement learning, and the medial prefrontal cortex. PsyArXiv. Published online September 4, 2023. https://doi.org/10.31234/osf.io/spxq9.
86. Gilboa, A. and Marlatte, H. (2017) Neurobiology of schemas and schema-mediated memory. *Trends Cogn. Sci.* 21, 618–631
87. Sadacca, B.F. *et al.* (2018) Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *eLife* 7, e30373
88. Tolman, E.C. and Honzik, C.H. (1930) Introduction and removal of reward, and maze performance in rats. *Univ. Calif. Publ. Psychol.* 4, 257–275
89. Tolman, E.C. (1948) Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208
90. Behrens, T.E.J. *et al.* (2018) What is a cognitive map? Organizing knowledge for flexiblebehavior. *Neuron* 100, 490–509
91. Schuck, N.W. *et al.* (2018) A state representation for reinforcement learning and decision-making in the orbitofrontal cortex. In *Goal-Directed Decision Making* (Morris, R. *et al.*, eds), pp. 259–278, Academic Press
92. Wikenheiser, A.M. and Schoenbaum, G. (2016) Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nat. Rev. Neurosci.* 17, 513–523
93. Garvert, M.M. *et al.* (2023) Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nat. Neurosci.* 26, 615–626
94. Wikenheiser, A.M. *et al.* (2017) Suppression of ventral hippocampal output impairs integrated orbitofrontal encoding of task structure. *Neuron* 95, 1197–1207.e3
95. Kaplan, R. *et al.* (2017) The role of mental maps in decision-making. *Trends Neurosci.* 40, 256–259
96. Eppinger, B. *et al.* (2023) Diminished state space theory of human aging. *Perspect. Psychol. Sci.* 17456916231204811,
97. Whittington, J.C.R. *et al.* (2022) How to build a cognitive map. *Nat. Neurosci.* 25, 1257–1272
98. Lopatina, N. *et al.* (2015) Lateral orbitofrontal neurons acquire responses to upshifted, downshifted, or blocked cues during unblocking. *eLife* 4, e11299
99. Lopatina, N. *et al.* (2017) Ensembles in medial and lateral orbitofrontal cortex construct cognitive maps emphasizing different features of the behavioral landscape. *Behav. Neurosci.* 131, 201
100. Mehta, P.S. *et al.* (2019) Ventromedial prefrontal cortex tracks multiple environmental variables during search. *J. Neurosci.* 39, 5336–5350
101. Rigotti, M. *et al.* (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590
102. Stoll, F.M. and Rudebeck, P.H. (2024) Preferences reveal dissociable encoding across prefrontal-limbic circuits. *Neuron* 112, 2241
103. Yoo, S.B.M. *et al.* (2018) Robust encoding of spatial information in orbitofrontal cortex and striatum. *J. Cogn. Neurosci.* 30, 898–913
104. Becket Ebitz, R. *et al.* (2020) Rules warp feature encoding in decision-making circuits. *PLoS Biol.* 18, e3000951
105. Fusi, S. *et al.* (2016) Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* 37, 66–74
106. Tye, K.M. *et al.* (2024) Mixed selectivity: cellular computations for complexity. *Neuron* 112, 2289–2303
107. Yoo, S.B.M. and Hayden, B.Y. (2020) The transition from evaluation to selection involves neural subspace reorganization in core reward regions. *Neuron* 105, 712–724.e4
108. Kennerley, S.W. *et al.* (2009) Neurons in the frontal lobe encode the value of multiple decision variables. *J. Cogn. Neurosci.* 21, 1162–1178
109. Botvinick, M. *et al.* (2020) Deep reinforcement learning and its neuroscientific implications. *Neuron* 107, 603–616
110. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533
111. Silver, D. *et al.* (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 1140–1144
112. Kiran, B.R. *et al.* (2022) Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans. Intell. Transp. Syst.* 23, 4909–4926
113. Bengio, Y. *et al.* (2014) representation learning: a review and new perspectives. *arXiv*, Published online April 24, 2014. http://doi.org/10.48550/arxiv.1206.5538
114. Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489
115. Heess, N. *et al.* (2017) Emergence of locomotion behaviours in rich environments. *arXiv*, Published online July 10, 2017. https://doi.org/10.48550/arXiv.1707.02286
116. Song, H.F. *et al.* (2017) Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* 6, e21492
117. Wierda, T. *et al.* (2023) Diverse and flexible behavioral strategies arise in recurrent neural networks trained on multisensory decision making. *bioRxiv*, Published online November 1, 2023. https://doi.org/10.1101/2023.10.28.564511v1

118. Zhang, Z. *et al.* (2018) A neural network model for the orbitofrontal cortex and task space acquisition during reinforcement learning. *PLoS Comput. Biol.* 14, e1005925
119. Dabney, W. *et al.* (2020) A distributional code for value in dopamine-based reinforcement learning. *Nature* 577, 671–675
120. Kozma, R. *et al.* (2018) Evolution of abstraction across layers in deep learning neural networks. *Procedia Comput. Sci.* 144, 203–213
121. Pessiglione, M. and Daunizeau, J. (2021) Bridging across functional models: the OFC as a value-making neural network. *Behav. Neurosci.* 135, 277
122. Cross, L. *et al.* (2021) Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron* 109, 724–738.e7
123. McGrath, T. *et al.* (2022) Acquisition of chess knowledge in AlphaZero. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2206625119
124. Silver, D. *et al.* (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv*, Published online December 5, 2017. http://arxiv.org/abs/1712.01815
125. Suhaimi, A. *et al.* (2022) Representation learning in the artificial and biological neural networks underlying sensorimotor integration. *Sci. Adv.* 8, eabn0984
126. Hattori, R. *et al.* (2023) Meta-reinforcement learning via orbitofrontal cortex. *Nat. Neurosci.* 26, 2182–2191
127. Wang, J.X. *et al.* (2018) Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* 21, 860–868
128. de Bruin, T. *et al.* (2018) Integrating state representation learning into deep reinforcement learning. *IEEE Robot. Autom. Lett.* 3, 1394–1401
129. Lesort, Timothée *et al.* (2018) State representation learning for control: an overview. *Neural Netw.* 108, 379–392
130. Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *Science* 313, 504–507
131. Anand, A. *et al.* (2020) Unsupervised state representation learning in Atari. *arXiv*, Published online November 5, 2020. https://doi.org/10.48550/arXiv.1906.08226
132. Banino, A. *et al.* (2018) Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 429–433
133. Fang, C. and Stachenfeld, K.L. (2023) Predictive auxiliary objectives in deep RL mimic learning in the brain. *arXiv*, Published online December 8, 2023. https://doi.org/10.48550/arXiv.2310.06089
134. Sandbrink, K. and Summerfield, C. (2024) Modelling cognitive flexibility with deep neural networks. *Curr. Opin. Behav. Sci.* 57, 101361
135. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253
136. Duan, Y. *et al.* (2016) Rl^2: fast reinforcement learning via slow reinforcement learning. *arXiv*, Published online November 10, 2016. https://doi.org/10.48550/arXiv.1611.02779
137. Sharpe, M.J. *et al.* (2019) An integrated model of action selection: distinct modes of cortical control of striatal decision making. *Annu. Rev. Psychol.* 70, 53–76
138. Wittkuhn, L. *et al.* (2021) Replay in minds and machines. *Neurosci. Biobehav. Rev.* 129, 367–388
139. Schuck, N.W. *et al.* (2015) Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron* 86, 331–340
140. De Martino, B. and Cortese, A. (2023) Goals, usefulness and abstraction in value-based choice. *Trends Cogn. Sci.* 27, 65–80
141. Molinaro, G. and Collins, A.G.E. (2023) A goal-centric outlook on learning. *Trends Cogn. Sci.* 27, 1150–1164
142. Noonan, M.A.P. *et al.* (2017) Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision- making in humans. *J. Neurosci.* 37, 7023–7035
143. Mack, M.L. *et al.* (2020) Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* 11, 46
144. Saxe, A.M. *et al.* (2018) *On the information bottleneck theory of deep learning*. International Conference on Learning Representation
145. Flesch, T. *et al.* (2022) Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* 110, 1258–1270.e11
146. Gaschler, R. *et al.* (2019) Incidental covariation learning leading to strategy change. *PLoS One* 14, e0210597
147. Löwe, A.T. *et al.* (2024) Abrupt and spontaneous strategy switches emerge in simple regularised neural networks. *PLoS Comput. Biol.* 20, e1012505
148. Loewe, A.T. *et al.* (2024) N2 sleep inspires insight. *bioRxiv*, Published online June 28, 2024. https://doi.org/10.1101/2024.06.24.600359
149. Mante, V. *et al.* (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78–84
150. Takagi, Y. *et al.* (2021) Adapting non-invasive human recordings along multiple task-axes shows unfolding of spontaneous and over-trained choice. *eLife* 10, e60988
151. Levy, D.J. and Glimcher, P.W. (2011) Comparing apples and oranges: using reward-specific and reward- general subjective value representation in the brain. *J. Neurosci.* 31, 14693–14707
152. Kuperwajs, I. *et al.* (2023) Using deep neural networks as a guide for modeling human planning. *Sci. Rep.* 13, 20269
153. Saxe, A.M. *et al.* (2019) A mathematical theory of semantic development in deep neural networks. *Proce. Natl. Acad. Sci. U. S. A.* 116, 11537–11546
154. Flesch, T. *et al.* (2018) Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci.* 115, E10313–E10322
155. Bellemare, M.G. *et al.* (2017) A distributional perspective on reinforcement learning. *arXiv*, Published online July 21, 2017. https://doi.org/10.48550/arXiv.1707.06887
156. Schultz, W. (1998) Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27
157. Muller, T.H. *et al.* (2024) Distributional reinforcement learning in prefrontal cortex. *Nat. Neurosci.* 27, 403–408